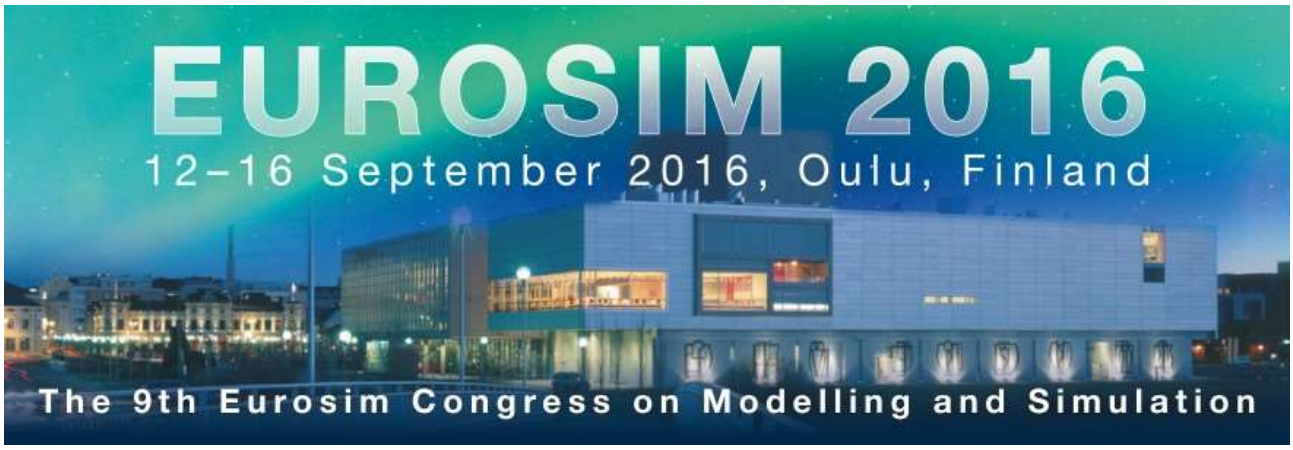
PROCEEDINGS OF

The 9th EUROSIM Congress on Modelling and Simulation EUROSIM 2016

The 57th SIMS Conference on Simulation and Modelling SIMS 2016



Editors: Esko Juuso, Erik Dahlquist, and Kauko Leiviskä

Organized by Scandinavian simulation society (SIMS),
Finnish Society
of Automation (FSA),
Finnish Simulation Forum (FinSim),
University of Oulu
and
Finnish Automation Support Ltd.

DOI: 10.3384/ecp17142

Proceedings of
The 9th EUROSIM Congress on Modelling and Simulation
EUROSIM 2016
and
The 57th SIMS Conference on Simulation and Modelling
SIMS 2016

Editors:
Esko Juuso, Erik Dahlquist and Kauko Leiviskä

Published by:
Scandinavian Simulation Society and Linköping University Electronic Press
ISBN: 978-91-7685-399-3
Series: Linköping Electronic Conference Proceedings, No. 142
ISSN: 1650-3686
eISSN: 1650-3740
DOI: 10.3384/ecp17142

Organized by:
Scandinavian simulation society (SIMS),
Finnish Society of Automation (FSA),
Finnish Simulation Forum (FinSim),
University of Oulu
and Finnish Automation Support Ltd.

Technically co-sponsored by:
IEEE Finland Section,
IEEE Computer Society,
International Federation of Automatic Control Technical Committees

Scandinavian Simulation Society
c/o Esko Juuso, Control Engineering, Faculty of Technology
P.O Box 4300
FIN-90014 University of Oulu
Finland

Copyright © Scandinavian Simulation Society, 2018

Preface

The 9th Eurosim Congress on Modelling and Simulation (Eurosim 2016) and the 57th SIMS conference on Simulation and Modelling (SIMS 2016) were held jointly in the City of Oulu, Finland. The Federation of European Simulation Societies (Eurosim) was set up in 1989. The purpose of EUROSIM is to provide a European forum for regional and national simulation societies to promote the advancement of modelling and simulation in industry, research and development. The program committee has organized an exciting and balanced program comprising presentations from distinguished experts in the field, and important and wide-ranging contributions on state-of-the-art research that provide new insights into the latest innovations in the field of modelling and simulation. EUROSIM 2016 was organized by Scandinavian simulation society (SIMS), Finnish Society of Automation (FSA), Finnish Simulation Forum (FinSim), University of Oulu and Finnish Automation Support Ltd. SIMS was founded in 1959. The event was technically co-sponsored by IEEE Finland Section, IEEE Computer Society and International Federation of Automatic Control Technical Committees on

- TC 3.2. Computational Intelligence in Control,
- TC 6.1. Chemical Process Control,
- TC 6.2. Mining, Mineral and Metal Processing,
- TC 6.3. Power and Energy Systems, and
- TC 6.4. Fault Detection, Supervision & Safety of Technical Processes-SAFEPROCESS.

Networking events were organized on three levels: Federation of European Simulation Societies (Eurosim) had Executive Board and Board meetings; Scandinavian simulation society (SIMS) had Board and Annual Meetings and Finnish Simulation Forum (FinSim), founded in 2001, had Board and Annual Meetings.

The overall management was done by the Eurosim Board, consisting of representatives of 17 European Simulation Societies lead by the Eurosim President Dr. Esko Juuso. The NOC was chaired by Prof. Kauko Leiviskä with Co-Chair Dr. Esko Juuso and Vice-Chair Industry Dr. Timo Ahola (FinSim). The IPC Chair Prof. Erik Dahlquist (SIMS), Co-Chair Prof. Bernt Lie (SIMS) and Vice-Chair Industry Dr. Lasse Eriksson were leading the IPC which consisted of 49 scientists. The Publication committee was chaired by Dr. Peter Ylén with C-Chairs Prof. Lars Eriksson and Prof. David Al-Dabass. Industry-based people participated in the congress committees. Eurosim and SIMS Boards were in the IPC. Eurosim members participated actively: 20 Eurosim countries from 14 member societies participated in the congress.

EUROSIM 2016 proved to be very popular and received submissions with authors from more than 40 countries. The conference program committee had a very challenging task of choosing high quality submissions. Each paper was peer reviewed by several independent referees of the program committee and, based on the recommendation of the reviewers, 184 regular papers were accepted for presentation. The papers offer stimulating insights into emerging modelling and simulation techniques and their applications in a wide variety of fields within science, technology, business, management and industry. Contributions are structured by

- Application domains, including Bio- and ecological systems, Building and construction, Economic and social systems, Energy, Industrial processes, Security and military, Transportation and vehicle systems, Water and wastewater, Weather and climate;
- Functionalities, including Control and optimization, Communication and security, Education and training, e-Learning, Fault detection & fault tolerant systems, Human-Machine interaction, Mechatronics and robotics, Planning and scheduling, Sensing, Virtual reality and visualization;
- Methodologies, including Computational intelligence, Conceptual modelling, Complex systems, Data analysis, Discrete event simulation, Distributed parameter systems, Parallel and distributed interactive systems, Simulation tools and platforms.
- Minisymposia integrate these areas in Control education, Wastewater, Applied energy, Solar thermal power plants, Complex dynamic systems, Chemical processes, Big data analysis and Soft computing, Innovative technology and Cooperative automation.

DOI: 10.3384/ecp17142

Panel discussions were organised for three topics: Modelling and Simulation in Processes and Cleantech, Future energy systems and Intelligent Systems and IoT in Future Automation. Three technical tours covered energy production, radio technology, printed electronics and mining. Tutorials were arranged on simulation tools and platforms.

The congress was very international: there were 181 participants from 33 countries. There were participants from all the continents, except Africa, but there were co-authors also from Africa. The program included 6 plenaries, 175 regular oral and 7 poster presentations. The focus was in Europe: 1/4 from Finland, 1/5 from other Nordic countries and 1/3 from other European countries. Far East was active: 26 participants, where 18 from Japan. In addition, there were 13 participants from other areas (Middle East, Americas and Australia). In total 22% of the participants were from outside Europe, almost as many as from Finland. The Proceedings includes includes 165 selected and revised papers.

Education for students was covered by the Control Education minisymposium organized The Education Committee of The Finnish Society of Automation and extended with papers from Education and Training, e-Learning track. IEEE Finland Section organized a workshop for early career development in the world of technology for young professionals. High school students visited some sessions.

We would like to express our sincere thanks to the plenary speakers, authors, session chairs, members of the program committee and additional reviewers who made this conference such an outstanding success. Finally, we hope that you will find the proceedings to be a valuable resource in your professional, research, and educational activities whether you are a student, academic, researcher, or a practicing professional.

Esko Juuso, Erik Dahlquist and Kauko Leiviskä

SIMS - Scandinavian Simulation Society



Suomen Automaatioseura ry
Finnish Society of Automation


Table of Contents

Congress location

The main venue and the exhibition site was the Oulu City Theatre in the City of Oulu, Capital of Northern Scandinavia.

Opening session, 13 September 2016

The Vice Rector for Cooperation affairs, Dr. Matti Sarén, University of Oulu, Finland
President of EUROSIM, Dr. Esko Juuso, University of Oulu, Finland
President of SIMS, Prof. Erik Dahlquist, Mälardalen University, Sweden
Congress Chair, Prof. Kauko Leiviskä, University of Oulu, Finland

Plenary presentations, 13 - 15 September 2016

Thermal Management Simulations within Power Engineering at ABB
Prof. Rebei Bel Fdhila, ABB Corporate Research, Sweden

Modelling and Simulation of the Electric Arc Furnace Processes
Assistant Professor Vito Logar, Laboratory of Modelling, Simulation and Control, Faculty of Electrical Engineering, University of Ljubjana, Slovenia

Part 1: Autonomous Driving and Levels of Automation,
Part 2: Situation Awareness and Early Recognition of Traffic Maneuvers
Dr Galia Weidl, Daimler AG, Germany

Simulating the Composition of the Atmosphere
Adjunct Professor Harri Kokkola, Atmospheric Research Centre of Eastern Finland, Finnish Meteorological Institute, Finland.

Using the Power of Simulation to bring Bottom Line Benefits to the Mining, Minerals and Metals Operations
Roy Calder, Director, Technical Sales, Global Solutions SimSci by Schneider Electric, Warrington, United Kingdom

Online Simulation Platform for Biophotonic Applications
Dr Alexey Popov, Optoelectronics and Measurement Techniques Laboratory, University of Oulu, Oulu, Finland

Congress topics, 13 - 15 September 2016

Application domains (pp. 19-340)

Functionalities (pp. 321-502)

Modelling and simulation methodologies (pp. 505-811)

Minisymposia, 13 - 15 September 2016 (pp. 812-1128)

Best practices and new trends in control education
Organizer: Finnish Society of Automation, Education Committee, Chair: Dr. Kai Zenger, Aalto University, Espoo, Finland

Modelling and control aspects in wastewater treatment processes
Organizer: Dr. Jesus Zambrano, Mälardalen University, Västerås, Sweden

Modelling and simulation in applied energy
Organizer: Prof. Erik Dahlquist, Malardalen University, Sweden

Modelling and simulation in solar thermal power plants
Organizer: Dr. Luis J. Yebra, CIEMAT, Plataforma Solar de Almería, Spain

Object-Oriented technologies of computer modelling and simulation of complex dynamical systems
Organizer: Russian Federation the National Simulation Society, Chair: Prof. Yuri Senichenkov St. Petersburg Polytechnic University, Russia

Chemical Process Systems Simulation
Organizer: Dr. Esko Juuso, University of Oulu, Finland

Industrial Optimization Based on Big Data Technology and Soft Computing
Organizers: Prof. Yukinori Suzuki, Muroran Institute of Technology, Japan, and Dr. Kai Zenger, Aalto University, Espoo, Finland

Simulation as Enabler for Innovative Technology
Organizer: Prof. Agostino G. Bruzzone, Genoa University, Italy

Cooperative Automation
Organizer: Dr. Esko Juuso, University of Oulu, Finland

Panel discussions, 13 - 15 September 2016

Modelling and Simulation in Cleantech
Chairs: Bernt Lie, University College of Southeast Norway, Norway
Jesús Zambrano, Mälardalen University, Sweden

Future Energy Systems
Chairs: Erik Dahlquist, Mälardalen University, Sweden,
Cristian Nichita. University of Le Havre, France

Intelligent Systems and IoT in Future Automation
Chairs: Esko K. Juuso, University of Oulu, Finland,
Lars Eriksson, Linköping University, Sweden

Workshops and tutorials

Career development in the world of technology and Young Professionals Meetup event
Chair: Rafal Sliz, University of Oulu, IEEE Finland Section

The basis of object-oriented modelling with Rand Model Designer
Prof. Yuri Senichenkov, St. Petersburg Polytechnic University, Russia

Introduction to modeling and simulation with Modelica using OpenModelica
Dr. Adrian Pop, Linköping University, Sweden

Closing session, 15 September 2016

Final Proceedings and Publication of revised papers
President of EUROSIM 2013-2016, Dr. Esko Juuso, University of Oulu, Finland

The Presidency period 2016 – 2019, EUROSIM 2019 Congress & Venue
President of EUROSIM 2016-2019, Prof. Emilio Jiménez Macías, CEA-SMSG, Spain

Technical tours, 16 September 2016

Tour I: Energia
Oulun Energia is the leading energy company in Northern Finland. Two power plants are visited:

- Laanila Eco Power Plant, 50 MW Waste to Energy Plant, uses municipal solid waste as a fuel.
- Toppila CHP Power Plants, 267 MW and 315 MW units, uses peat and wood as a fuel.

Tour II: Radio Technology and Printed Intelligence

- Factory tour at Nokia's Radio Technology Center Site.
- VTT, PrintoCent Pilot Factories, including the new 5G Test Environment and Maxi IoT pilot line

Tour III: Mining Engineering
Oulu Mining School (OMS) integrates scientific disciplines along the value chain from exploration to mining and mineral processing. OMS has a unique continuous and automated concentrator minipilot facility.

Social program

Welcoming reception, The Oulu City Theatre, 12 September 2016.

The City-Reception, Oulu City Hall, 13 September 2016.

Banquet, Hotel Radisson Blu Oulu, 14 September 2016

Organization

EUROSIM President
Dr. Esko Juuso, *University of Oulu, Finland*

Congress Chair
Prof. Kauko Leiviskä, University of Oulu, Finland

Congress Vice-Chair, Industry
Dr. Timo Ahola, FinSim, Outokumpu Stainless Ltd, Finland

Programme Chair
Prof. Erik Dahlquist, SIMS, Mälardalen University, Sweden

Programme Co-Chair
Prof. Bernt Lie, SIMS, University College of Southeast Norway, Norway

Programme Vice- Chair, Industry
Dr. Lasse Eriksson, FSA, Cargotec Corporation, Finland

Exhibition Chair
Dr. Pekka Tervonen, University of Oulu, Finland

Publication Chair
Dr. Peter Ylén, FSA, VTT Ltd, Finland

Publication Co- Chair
Prof. Lars Eriksson, SIMS, Linköping University, Sweden

Publication Co-Chair, IEEE-CPS
Prof. David Al-Dabass, UKSim, Nottingham Trent University, United Kingdom

EUROSIM Board Members

Esko Juuso	President	Finland
Borut Zupančič	Secretary, SLOSIM repr.	Slovenia
Felix Breitenecker	Treassurer, ASIM repr.	Austria
Khalid Al-Begain	Past president	UK
Emilio Jimenez Macías	CEA-SMSG repr.	Spain
Vesna Dušak	CROSSIM repr.	Croatia
Miroslav Šnorek	CSSS repr.	Czech Republic
Miguel Mujica Mota	DBSS repr.	Netherlands
Karim Djouani	FRANCOSIM repr.	France
Gábor Szűcs	HSS repr.	Hungary
Franco Maceri	ISCS repr.	Italy
Agostino Bruzzone	Leophant Repr.	Italy
Yuri Merkuryev	LSS repr.	Latvia
Tadeusz Nowicki	PSCS repr.	Poland
Yuri Senichenkov	NSSM	Russian Federation
Erik Dahlquist	SIMS	Sweden
Alessandra Orsoni	UKSim	UK
Henri Pierreval	Publications	France

National Organizing Committee (NOC)

Prof. Kauko Leiviskä, University of Oulu, Chair
Dr. Esko Juuso, University of Oulu, EUROSIM President, Co-Chair
Dr. Timo Ahola, Outokumpu Stainless Ltd, FinSim Chair, Vice-Chair, Industry
Dr. Peter Ylén, VTT Ltd, Publication Chair
Dr. Pekka Tervonen, University of Oulu, Exhibition Chair
Dr. Rafal Sliz, University of Oulu, IEEE Finland Section
Dr. Ari Isokangas, University of Oulu
Mr. Marko Vuorio, Finnish Automation Support Ltd
Ms. Anu Randén-Siippainen, Finnish Automation Support Ltd

International Program Committee

Prof. Erik Dahlquist, SIMS, Sweden, Chair

Prof. Bernt Lie, SIMS, Norway, Co-Chair

Dr. Lasse Eriksson, FSA, Finland, Vice-Chair, Industry

Dr. Timo Ahola, FinSim, Finland

Prof. Kahlid Al-Begain, EUROSIM Past President, UKSim, U.K.

Prof. David Al-Dabass, UKSim, U.K.

Prof. Mikulas Alexik, EUROSIM Past President,CSSS, Slovakia

Prof. Leon Bobrowski, PSCS, Poland

Prof. Felix Breitenecker, ASIM, Austria

Prof. Agostino Bruzzone, Liophant, Italy

Dr. Tero Eklin, Finland

Prof. Brian Elmegaard, SIMS, Denmark

Prof. Lars Eriksson, SIMS, Sweden

Prof. Timo Fabritius, Finland

Prof. Peter Fritzson, SIMS, Sweden

Prof. Diego Galar, Sweden

Prof. Len Gelman, U.K.

Prof. Edmond Hajrizi, KA-SIM, Kosovo

Markku Henttu, FinSim, Finland

Prof. Yrjö Hiltunen, Finland

Prof. Emilio Jiménez Macías, CEA-SMSG, Spain

Dr. Axel Ohrt Johansen, SIMS, Denmark

Prof. Magnus Jonsson, SIMS, Iceland

Dr. Kaj Juslin, EUROSIM Past President, SIMS, Finland

Dr. Esko Juuso, EUROSIM President, Finland

Prof. Mika Järvinen, Finland

Dr. Marko Kesti, Finland

Dr. Jónas Ketilsson, SIMS, Iceland

Prof. Tiina Komulainen, SIMS, Norway

Timo Korpela, FinSim, Finland

Prof. Andreas Kugi, ASIM, IFAC, Austria

Prof. Kauko Leiviskä, Finland

Dr. Mika Liukkonen, Finland

Prof. Seppo Louhenkilpi, Finland

Yrjö Majanne, FinSim, Finland

Dr. Toni Mattila, IEEE Finland Section

Prof. Yuri Merkurjev, LSS, Latvia

Prof. Tadeusz Nowicki, PSCS, Poland

Prof. Gerhard Nygaard, SIMS, Norway

Dr. Alessandra Orsoni, UKSim, UK

Dr. Marko Paavola, Finland

Prof. Henri Pierreval, France

Dr. Jari Ruuska, FinSim, Finland

Dr. Mika Ruusunen, Finland

Prof. Yuri Senichenkov, NSSM, Russia

Dr. Rafal Sliz, IEEE Finland Section

Prof. Miroslav Šnorek, CSSS, the Czech Republic

Dr. Aki Sorsa, Finland

Prof. Kim Sørensen, SIMS, Denmark

Dr. Luis J. Yebra, Spain

Dr. Peter Ylén, FSA, Finland

Dr. Kai Zenger, FSA, Finland

Prof. Borut Zupančič, EUROSIM Secretary, EUROSIM Past President, SLOSIM, Slovenia

International Reviewers.

Title	Givenname	Surname	Affiliation	Country
Dr.	Ghulam	Abbas	GIK Institute of Engineering Sciences & Technology	Pakistan
Ms.	Azian Azamimi	Abdullah	Nara Institute of Science and Technology	Japan
Prof.	Majida	Alasady	University of Tikrit	Iraq
Mr.	Boon Chong	Ang	Intel	Malaysia
Mr.	Peyman	Arebi	Technical and Vocational University - Technical and Vocational College of Boushehr	Iran
Prof.	Eduard	Babulak	The Institute of Technology and Business in Ceske Budejovice	Czech Republic
Dr.	Kambiz	Badie	Iran Telecom Research Center	Iran
Dr.	Arijit	Bhattacharya	University of Dubai	United Arab Emirates
Dr.	Muhammad H. F.	Bin Md Fauadi	Universiti Teknikal Malaysia Melaka	Malaysia
Dr.	Silvia	Cateni	Scuola Superiore Sant'Anna, Pisa	Italy
Prof.	Sung-Bae	Cho	Yonsei University	Korea
Mr.	G	Deka	Directorate General of Training	India
Mr.	Hosam	Faiq	Universiti Sains Malaysia	Malaysia
Mr.	Jitender	Grover	Maharishi Markandeshwar University	India
Mr.	Petri	Hietaharju	University of Oulu	Finland
Dr.	Ari	Isokangas	University of Oulu	Finland
Prof.	Sudhanshu	Jamuar	University Malayasia Perlis	Malaysia
Dr.	Dayang	Jawawi	Universiti Teknologi Malaysia	Malaysia
Dr.	Kponyo	Jerry	Kwame Nkrumah University of Science and Technology	Ghana
Prof.	Rihard	Karba	University of Ljubljana	Slovenia
Prof.	Tommi	Karhela	Aalto University	Finland
Mr.	Konsta	Karioja	University of Oulu	Finland
Prof.	S. D.	Katebi	Shiraz University, Shiraz	Iran
Prof.	Dong-hwa	Kim	Hanbat National University	Korea
Prof.	Miroljub	Kljajic	University of Maribor	Slovenia
Mr.	Antti	Koistinen	University of Oulu	Finland
Dr.	J. Mailen	Kootsey	Simulation Resources, Inc.	USA
Dr.	Binod	Kumar	JSPM's Jayawant Institute of Computer Applications, Pune	India
Mr.	Jouni	Laurila	University of Oulu	Finland
Dr.	Toni	Liedes	University of Oulu	Finland
Mr.	Solomon	Mangeni	Swansea University	United Kingdom
Prof.	Rashid	Mehmood	King AbdulAziz University	Saudi Arabia
Prof.	Gasper	Music	University of Ljubljana	Slovenia
Prof.	Britt	Moldestad	University of South-Eastern University	Norway
Ms.	Outi	Mäyrä	University of Oulu	Finland
Dr.	Raza	Naqvi	Mälardalen University	Sweden
Dr.	Sophan W.	Nawawi	Universiti Teknologi Malaysia	Malaysia
Mr.	Riku-Pekka	Nikula	University of Oulu	Finland
Dr.	Kenneth	Nwizege	Ken Saro-Wiwa Polytechnic, Bori	Nigeria
Dr.	Markku	Ohenoja	University of Oulu	Finland

Dr.	Olugbenga	Olubodun	University of Swansea	United Kingdom
Dr.	Alessandra	Orsoni	Kingston University	United Kingdom
Dr.	Mirjana	Pejic-Bach	University of Zagreb	Croatia
Dr.	Danilo	Pelusi	University of Teramo	Italy
Mr.	Mika	Pylvänäinen	University of Oulu	Finland
Dr.	Ahmad	Jordehi	University Putra Malaysia	Malaysia
Prof.	Philip	Sallis	Auckland University of Technology, NZ	New Zealand
Mr.	Vaclav	Satek	Brno University of Technology	Czech Republic
Prof.	Zaliman	Sauli	Universiti Malaysia Perlis	Malaysia
Dr.	Vivek	Sehgal	Jaypee University of Information Technology	India
Dr.	Silja	Sigurdardottir	Reykjavik University	Iceland
Dr.	Shobhana	Singh	Aalborg University	Denmark
Dr.	Suchitra	Sueeprasan	Chulalongkorn University	Thailand
Mr.	Mohamad Fani	Sulaima	Universiti Teknikal Malaysia Melaka	Malaysia
Mr.	Irfan	Syamsuddin	State Polytechnic of Ujung Pandang	Indonesia
Prof.	Geetam	Tomar	Machine Intelligence Research (MIR) Labs Gwalior	India
Mr.	Jani	Tomperi	University of Oulu	Finland
Dr.	Gancho	Vachkov	The University of the South Pacific (USP)	Fiji
Dr.	Patrick	Wang	Northeastern University	USA
Dr.	Peter	Ylen	VTT Technical Research Centre of Finland	Finland
Dr.	Jesus	Zambrano	Mälardalen University	Sweden
Dr.	Jovana	Zoroja	University of Zagreb	Croatia
Prof.	Lars	Øi	University of South-Eastern University	Norway

Publication Committee

Dr. Peter Ylén, VTT Ltd, Publication Chair
Prof. Lars Eriksson, Linköping University, Publication Co-Chair
Prof. David Al-Dabass, Nottingham Trent University, UK, Publication Co-Chair IEEE-CPS
Prof. Erik Dahlquist, Mälardalen University, SIMS President, IPC Chair
Prof. Bernt Lie, Telemark University, IPC Co-Chair
Dr. Esko Juuso site, University of Oulu, EUROSIM President
Dr. Rafal Sliz, University of Oulu, IEEE Finland Section
Dr. Jari Ruuska, University of Oulu

Exhibition Committee

Dr. Pekka Tervonen, University of Oulu, Chair
Dr. Harri Happonen, Fimlab Laboratories, Finland, Co-Chair
Dr. Esko Juuso, University of Oulu, EUROSIM President
Dr. Ari Isokangas, University of Oulu
Mr. Marko Vuorio, Finnish Automation Support Ltd

Plenary presentations

Plenary 1

Thermal Management Simulations within Power Engineering at ABB

Prof. Rebei Bel Fdhila

ABB Corporate Research
Sweden
Email: rebei.bel_fdhila@se.abb.com

The area of thermal management is driven by miniaturization in industry e.g. power electronics, motors or transformers. It is a natural response to size restrictions as in automotive and robotics, to space excessive costs e.g. offshore applications or simply to industrial or comfort requirements. Besides that, unceasing market and technology demand for higher currents, higher voltages or higher power is inevitably leading to a substantial power density increase justifying large losses and generating important amounts of heat. Confined electrical systems, enclosures containing electrical components and other apparatus and devices can generate a lot of heat able to significantly reduce the life time of an installation if no appropriate thermal solutions were adopted. Cooling is also needed to provide the appropriate process or product quality with minimizing energy consumption and environmental impact. An ever-increasing power density drives the need for more effective thermal management solutions where several phenomena e.g. electromagnetic, thermal and/or mechanical can be simultaneously taken into account. ABB is a leading company within power and automation technologies and to maintain its product quality and market penetration has always invested in acquiring state-of-the-art hardware and software tools to cope with its technology needs and ambitions. We are also building our own integrated multiphysics simulation methods able to develop accurate thermal solutions that account for the major interacting physical phenomena. This presentation can introduce you to our know-how in terms of numerical predictions of coupled systems and will also provide you with several examples of solutions where advanced simulations have been used.

Biography

Rebei Bel Fdhila (male), Adjunct Professor in Process Modelling and Computational Fluid Dynamics at Mälardalen University since 2006. Got a PhD in 1991 from the National Polytechnic Institute of Toulouse, France “INP Toulouse/ENSEEIHT” within multiphase flows and worked as a post-doc with EDF and CNRS in France followed by Twenty University in Holland. Since 1995 he joined ABB Corporate Research in Sweden first as a researcher and today acting in his global role as a Corporate Research Fellow in Thermal Management. He has a large experience within the advanced modeling and simulation world. 30+ publications and 9 active patent families.

Plenary 2

Modelling and Simulation of the Electric Arc Furnace Processes

Assistant Professor Vito Logar

Laboratory of Modelling, Simulation and Control,
Faculty of Electrical Engineering, University of Ljubjana, Slovenia

Email: vito.logar@fe.uni-lj.si

Current market demands on steel quality, price and production times dictate the introduction of several technological innovations regarding the electric arc furnace (EAF) steelmaking. One of the fields, which is rapidly developing and has significant potential is related to the advanced software support of the EAF operation, which combines data acquisition, advanced monitoring and proper control of the EAF. This paper briefly presents the idea and development of all key EAF-process models, which are together with measured EAF data used to estimate the unmeasured process values. The models are based on fundamental physical laws and are implemented mainly using nonlinear, time-variant ordinary differential equations. The validation results that were performed using operational EAF measurements indicate high levels of estimation accuracy and the final outcome of the study results in a fully operational EAF model, describing all crucial steel-recycling processes. The accuracy of the presented models is in the range of +/- 15 K for steel temperature and +/-10 % for steel composition. Therefore, the versatility and accuracy of the models allows the usage of the models in broader software environments in a form of soft sensors for process monitoring, process optimization and operator decision support.

Biography

Vito Logar is an Assistant Professor at the Faculty of Electrical Engineering, Univ. of Ljubljana. He is working on the described area for many years in several projects. So his research interests include modelling and optimization techniques regarding the electric arc furnace steel recycling processes. In 2013 he received the award for outstanding scientific achievement for the year 2011 from the Slovenian Research Agency (ARRS). In 2014 he received the award for outstanding scientific and pedagogic achievements from the University of Ljubljana. He is currently also the president of the Slovenian society for modelling and simulation SLOSIM.

More info on EAF modelling and simulation: EAF Simulator: http://msc.fe.uni-lj.si/eaf.asp
More info on the research: ResearchGate: https://www.researchgate.net/profile/Vito_Logar

Plenary 3

Situation Awareness and Early Recognition of Traffic Maneuvers

Dr Galia Weidl

Daimler AG, Germany

Email: galia_weidl@hotmail.com; galia.weidl@daimler.com

We outline the challenges of situation awareness with the early and accurate recognition of traffic maneuvers and how to assess them. This includes also an overview of the available data and derived situation features, handling of data uncertainties, modelling and the approach for maneuver recognition. An efficient and effective solution, meeting the automotive requirements, is successfully deployed and tested on a prototype car. Test driving results show that earlier recognition of intended maneuver is feasible on average 1 second (and up to 6.72 s) before the actual lane marking crossing. The even earlier maneuver recognition is dependent on the earlier recognition of surrounding vehicles.

Keywords – bayesian networks, massive data streams

Biography

Galia Weidl obtained the MSc.degree in physics and mathematics from St.Petersburg State University, Russia, in 1993, and Fil.Lic. degree in theoretical physics from the University of Stockholm, Sweden, in 1996, and a Tekn.Dr. doctoral degree in process engineering from Mälardalen University, Sweden in 2002. Until 2006 she held a postdoctoral appointment at Stuttgart University, Germany. She has held appointments with the research teams at ABB Sweden (1997-2002), Bosch (2006-2008) and Daimler (since 2008). Her current research topic focuses on Bayesian networks in the area of autonomous driving. Galia Weidl was appointed in June 2015 by the European Commission as invited independent expert for Horizon2020.

Plenary 4

Simulating the Composition of the Atmosphere

Adjunct Professor Harri Kokkola

Atmospheric Research Centre of Eastern, Finland
Finnish Meteorological Institute, Finland.

Email: harri.kokkola@fmi.fi

Climate models are an essential tool when estimating how climate will change in the future. The atmospheric core of these models simulates the circulation of the atmosphere by solving fundamental physical equations of conservation of motion, mass, and energy as well as the equation of state. However, climate is affected also by several other processes than the atmospheric circulation and to get an accurate projection of future climate, it is necessary to incorporate all these processes in the model. Such processes include cloud formation (warm clouds, ice clouds), cryosphere (ice/snow), land surface (soil, reflectance), biosphere (ecosystems, agriculture), ocean (heat transport). These processes are calculated with individual submodels which are coupled to the core atmospheric model and they are also coupled to each other so that they interact. However, machine learning methods and emulator techniques are emerging in the climate science. We have investigated the potential of these methods to decrease the error coming from simplifications of aerosol processes in global aerosol models. Our results show that machine learning methods can significantly increase the accuracy of coarse aerosol models without significantly increasing their computational burden.

Biography

Harri Kokkola is the group leader of Atmospheric Modeling group at the Research Centre of Eastern Finland, Finnish Meteorological Institute. He is working on atmospheric modeling and aerosol-cloud interactions. The main focus in his research is global scale aerosol-climate modeling and has been one of the main developers of the aerosol-chemistry-climate model ECHAMHAMMOZ. His research group has developed an aerosol microphysics module SALSA which has been implemented in a cloud scale model, an air quality model as well as regional and global climate models. They are also actively involved in AeroCom project which is an open international initiative of scientists interested in the advancement of the understanding of the global aerosol and its impact on climate.

Plenary 5

Using the Power of Simulation to bring Bottom Line Benefits to the Mining, Minerals and Metals Operations

Roy Calder

Director, Technical Sales, Global Solutions
SimSci by Schneider Electric, Warrington, United Kingdom

Email: roy.calder@schneider-electric.com

While today's economic climate for the Mining, Metals and Minerals (MMM) industries may not be at its best; the industry is still faced with the need to deliver products cost effectively, at the correct specification, while maintaining a high level of safety for the plant and the personnel. For many years the MMM industry has lagged behind the Hydrocarbon Processing Industry (HPI) with regard the use of Simulation, however as the level of investment has grown over the years the need to ensure cost effective design, allied to improvements in delivery time the use of Simulation has become an integral part of the design, construction and commissioning of new plants across the globe. One area Simulation applications are proving themselves is in the area of Operational Safety. The HPI has long used Simulation based Operator Training Simulators to ensure safe operations and this is being carried over with increasing uptake happening in the MMM industry. In the future as the MMM industry becomes increasingly sophisticated at the same time facing the difficulties of shrinking bottom lines it is clear that Simulation will become fundamental in delivering the tools and solutions that will enable the industry to ensure growth in its bottom line in years to come. Schneider-Electric has had a long history in the MMM industry and this paper will highlight how the acquisition of Invensys has brought a completely new perspective to the industry and, most importantly, allows MMM companies to grow their Bottom Line.

Biography

BSc in Chemical Engineering, 1984, University of Strathclyde, Glasgow, then wide ranging industrial experience as: 1) in South Africa as a Metallurgist on Westonaria Gold mine, 2) SASOL in Rosebank as a Process Engineer, 3) L'Air Liquide, 4) INHER SA as Divisional Head, Process Engineering in 1991, where he managed the SULZER Chemtech Agency delivering process plant in multiple industries, 5) Process Sales Director, BHR, UK, 6) SimSci Division of INVENSYS, now Schneider-Electric. Currently Director of Technical Sales, SimSci regional team, 80 strong, on Simulation in engineering and operations community of industries as varied as Oil Production, gold and coal mining and the power industry. The team directly serve the EURA (Europe, Russia and Africa) activities of SimSci. Written numerous papers and presented at World Petroleum Congress, ERTC, SAICHE, AICHE, IChemE & DECHEMA events as well as numerous industrial symposiums. He is joint holder of a patent on the application of Structured Packing in Wax Separation.
He is devoted to Rugby, though no longer playing, and heads the Mini and Junior Section of his local club of some 300 budding young players.

Plenary 6

Online Simulation Platform for Biophotonic Applications

Alexey Popov [1] , Alexander Bykov [1] , Alexander Doronin [2] , Hannu Sorvoja, and Igor Meglinski [1]

[1] Optoelectronics and Measurement Techniques Laboratory, University of Oulu, Oulu, Finland
[2] Computer Graphics Group, Department of Computer Science, Yale University, New Haven, USA

Email: popov@ee.oulu.fi

Currently optical methods are gaining ground for biomedical applications such as cancer and cardiovascular diagnostics, dermatology, ophthalmology, pharmaceutical research, cosmetics and healthcare industry. Benefits of optical techniques are their non-invasiveness, ability for remote monitoring and access to biological objects from cell to body level, to name a few.

The light irradiation dose, measurement volume, sensitivity of optical modalities are of crucial importance in biomedical diagnostics before implementing the developed techniques in in vivo research and clinical trials. An essential part of the preliminary studies is the use of phantoms and simulations for the optimal configuration of the setup and refining the measurement procedure. Up to now, such simulations were performed in every lab using own codes and local resources.

We report about the next step in the computational diagnostics, an online computational platform for the needs of biomedical optics and biophotonics. The platform serves as a tool for calculation of a sampling volume, fluence rate, skin spectrum, skin colour, and a number of optical techniques including optical coherence tomography, polarization, coherent backscattering, pulse oxymetry, confocal microscopy, fluorescence, and diffuse wave spectroscopy.

We used the inheritance feature of Object-oriented programming (OOP) to create a ‘smart’ hierarchical structure of the Monte Carlo (MC) code to avoid having multiple classes for similar tasks. The hierarchy allows ‘allied’ objects to share variables and members, significantly reducing the amount of source code and paving the way to extend and generalize the MC. Depending on the application, objects can be tuned to an appropriate state of light-tissue interaction and to a particular optical diagnostic technique.

To achieve optical simulation performance, we employed a developed parallel computing framework known as Compute Unified Device Architecture (CUDA), introduced by the NVIDIA Corporation. Specially designed for professional 3D graphics applications, this technology allows each graphic chip to be logically divided into hundreds of cores, turning the graphics processing unit into a massive co-processor for parallel computations. This capability enables the simulation of thousands of objects, i.e. the simultaneous propagation of photons in the medium - that speeds the process of simulation up to 103 times.

The computational solution utilizes recent developments in HyperText Markup Language (HTML) 5, accelerated by the graphics processing units (GPUs), and therefore is convenient for the practical use at the most of modern computer-based devices and operating systems. Figure 1 shows the interactive user interface for selecting a particular MC application. The results of imitation of human skin reflectance spectra and the corresponding skin colors are presented in Figure 2.

The platform can ease research in a number of areas and can be used for professional and educational purposes.

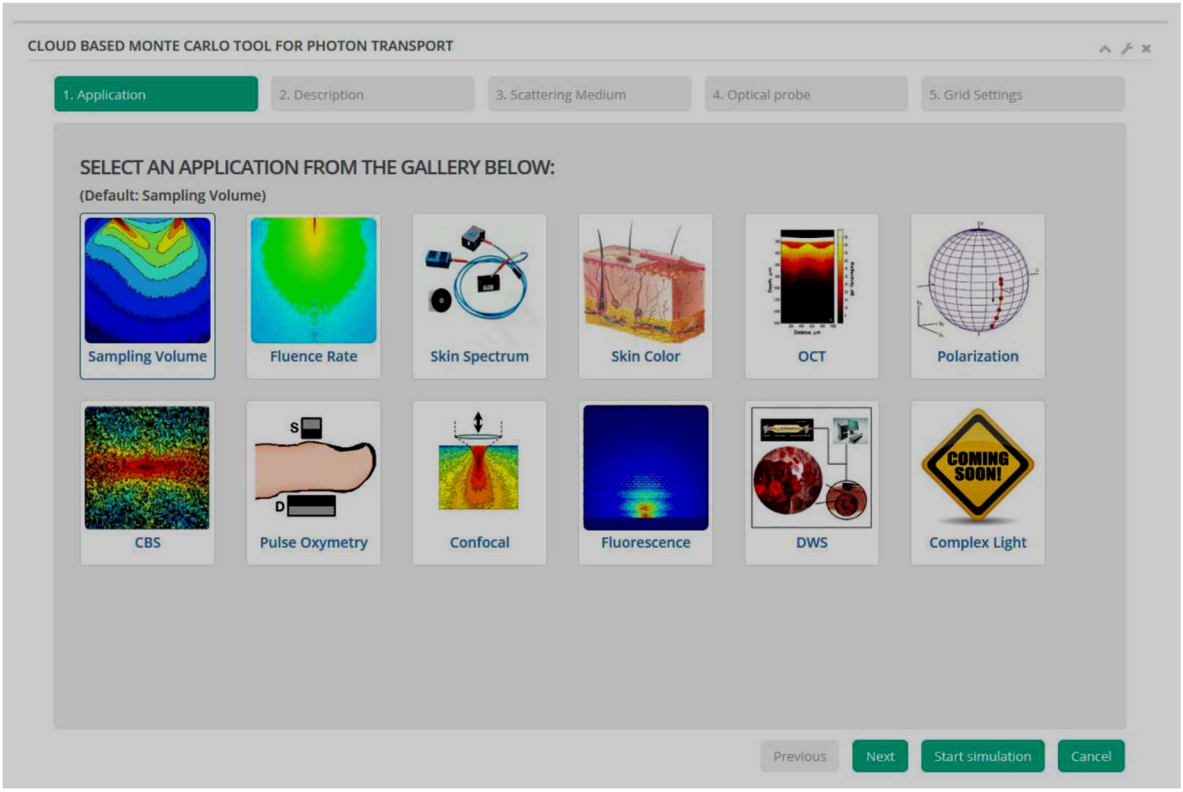


Figure 1. A variety of options offered by the online platform (www.biophotonics.fi).

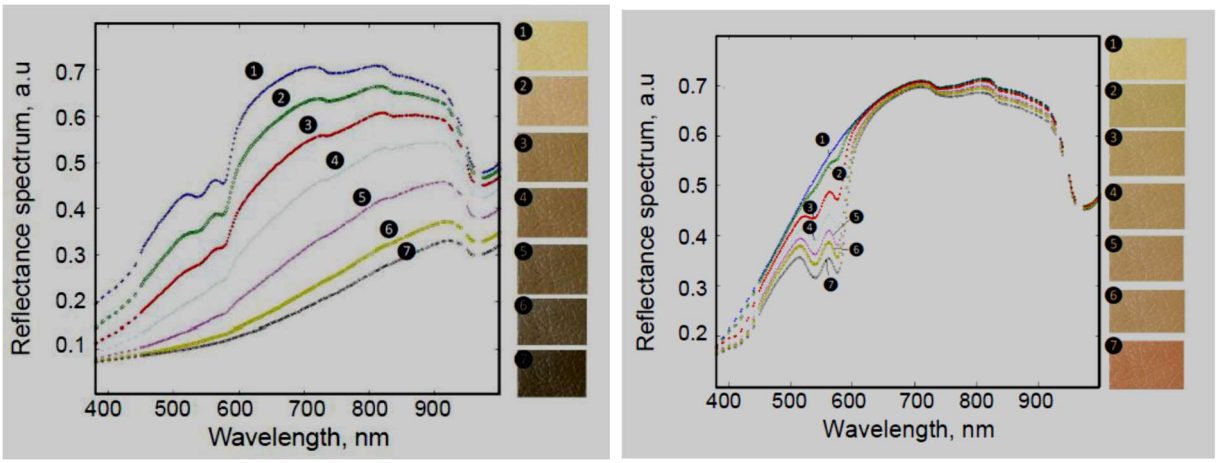


Figure 2. Results of MC simulations of human skin spectra and corresponsing colors while varying the melanin content in living epidermis (left): (1) – 0%, (2) – 2%, (3) – 5%, (4) – 10%, (5) – 20%, (6) – 35%, (7) – 45%; and while varying the blood concentration (right) in the layers from papillary dermis to subcutaneousl tissue: (1) – 0%, (2) – 2%, (3) – 5%, (4) – 10%, (5) – 20%, (6) – 35%, (7) – 45%, respectively.

References:

1. A. Doronin and I. Meglinski. Online object oriented Monte Carlo computational tool for the needs of biomedical optics, Biomed. Opt. Express 2(9): 2461-2469, 2011.
2. A. Doronin, H. Rushmeier, I. Meglinski, and A. Bykov. Cloud-based Monte Carlo modelling BSSRDF for the rendering of human skin appearance. In Proc. SPIE 9719, Biophysics, Biology, and Biophotonics: the Crossroads, 97190F, 2016. doi:10.1117/12.2214512.

Biographies of the authors

Alexey Popov, D.Sc. (Tech.) is a Senior Researcher and Docent in Optoelectronics and Measurement Techniques Laboratory at the University of Oulu, Finland. He graduated from the Physics Department of M.V. Lomonosov State University (Russia) with M.Sc. degree in 2003 and was awarded with PhD degree in 2006. He received his D.Sc. (Tech.) degree from the Faculty of Technology of the University of Oulu (Finland) in 2008. He is an author of 90 papers in international peer reviewed journals and SPIE proceedings and ca. 100 presentations at major international conferences, symposia and workshops including 15 invited lectures. Currently, he is a Senior Researcher and Docent in the Optoelectronics and Measurement Techniques Laboratory at the University of Oulu; a member of SPIE and a Faculty Advisor of the SPIE Student Chapter of the University of Oulu, Northernmost and 1st in Finland.

Alexander Bykov, Ph.D. (Phys.) and D.Sc. (Tech.), born in 1981, is curently a Postdoctoral Researcher at the Optoelectronics and Measurement Techniques Laboratory, University of Oulu. He has over ten-year experience in research in the fields of photonics and biomedical optics. He received M.Sc. diploma in Physics at the M.V. Lomonosov Moscow State University in 2005 and Ph.D. in 2008 from the same university. In 2010, he received D.Sc.(Tech.) degree from the Faculty of Technology at the University of Oulu and continued as a postdoctoral researcher at the Optoelectronics and Measurement techniques laboratory. He is an author and co-author of over 60 scientific papers published in refereed international journals and book chapters, cosupervisor of undergraduate and postgraduate students.

Alexander Doronin is a Postdoctoral Associate in Computer Science working in Computer Graphics Group, Yale University, USA. His research interests are interdisciplinary and lie at the interface between Computer Science, Physics, Optics and Biophotonics focusing on Physically-Based Rendering, Development of realistic material models, Monte Carlo modeling of light transport in turbid media, Color Perception, Translucency, Appearance and Biomedical Visualization.

Hannu Sorvoja, D.Sc. (Tech.), born in 1966, is currently a Laboratory Manager at the Optoelectronics and Measurement Techniques Unit, University of Oulu. He has over twenty-year experience in research in the fields of biomedical engineering. He received M.Sc.(Tech.) in Electrical Engineering 1993, Lic.Sc.(Tech.) in 1998, and D.Sc.(Tech.) 2006, all from the Faculty of Technology at the University of Oulu, and continued as a professor and a postdoctoral researcher. He is an author and co-author over 40 scientific papers published in refereed international journals or conferences and three patents. In addition, he has supervised over 30 M.Sc.(Tech.) and Lic.Sc.(Tech.) theses.

Professor Igor Meglinski, Ph.D. is Head of Optoelectronics and Measurement Techniques Laboratory, Faculty of Information Technology and Electrical Engineering, University of Oulu. He has over 20 years experience in biomedical optics, biomedical engineering, medical physics, and sensor technologies. He is an author and coauthor of over 200 research papers in the peer reviewed scientific journals, proceedings of international conferences and book chapters, and over 400 presentations at the major international conferences and symposia, including over 200 invited lectures and plenary talks. His research interests lie at the interface between physics, medicine, and biological sciences, focusing on the development of new non-invasive imaging/diagnostic techniques and their application in medicine and biology, material sciences, pharmacy, food, environmental monitoring, and health care industries. For the last ten years, he has been a Principal Investigator and/or Coordinator for over 60 research projects, supported by various funding bodies, including UK NHS trust, NATO, Royal Society, U.S. CRDF, New Zealand Ministry of Business, Innovation & Employment, Maurice Wilkins Centre (MWC), New Zealand Ministry of Foreign Affairs and Trade, A*STAR (Singapore), Federal Agency for Science and Innovations (Russia), Weizmann Institute of Science (Israel) and industrial partners including Procter & Gamble, Philips, General Electrics, Unilever and other (with a total cumulative budget of over $16M). Prof. Meglinski is a Fellow of the Institute of Physics (London, UK) and Fellow of SPIE.

Selected and revised papers

Keynotes

Track A01. Bio-/Ecological Systems: Agriculture, Bioinformatics/Bioengineering, Biological/Medical Systems

Track A02. Building/Construction: Automation, Engineering, Built Environment, Energy/Health

Track A03. Economic and Social Systems: Computational Finance/Economics, Control Education

Track A04. Energy Systems: Electricity/Heat/Gas Networks, Geothermal, Hydropower, Plants, Smart Grids for Heat/Electricity, Solar, Wind

Track A05. Industrial Processes: Chemical, Forest, Manufacturing, Metal, Mining/Mineral Processing, Pharmaceutical Industry

Track A06. Security and Military

Track A07. Transportation/Vehicle Systems, Aerospace/Automotive Applications, Autonomous Systems/Vehicles, Harbour/Shipping/Marine,Logistics, Vehicle Systems

Track A08. Water/Waste-water: Treatment Plants and Networks

Track A09. Other Application Domains

Track F01. Control and Optimization: Computers in Control, Adaptation, Intelligent Analyzers, Model-based Control

Track F02. Communication and Security: Internet/Cloud Computing, Wireless, Security

Track F03. Education and Training, e-Learning

Track F04. Fault Detection & Fault Tolerant Systems: Condition Monitoring, Maintenance

Track F05. Mechatronics and Robotics

Track F06. Planning and Scheduling

Track F07. Sensing: Image, Speech and Signal Processing. Circuits, Sensors and Devices

Track F08. Virtual Reality and Visualization, Computer Art, Serious Games, Visualization

Track M01. Computational Intelligence: Evolutionary, Fuzzy, Knowledge, Natural Language, Nature Inspired, Neural/Neuro-fuzzy, Patterns/Machine Intelligence

Track M02. Conceptual Modelling

Track M03. Complex Systems

Track M04. Data Analysis: Fractional Differentiation, Reinforcement Learning, Semantic Mining, Statistical Analysis

Track M05. Discrete Event Simulation

Track M06. Distributed Parameter Systems: Computational Fluid Dynamics, Partial Differential Equations, Stochastic Systems

Track M07. Parallel and Distributed Interactive Systems

Track M08. Simulation Tools/Platforms: Domain-Specific Tools, Simulation Software, Hardware in the Loop, Verification and Validation

Track M09. Other Methodologies

Track S01. Best Practices and New Trends in Control Education

Track S02. Modelling and Control Aspects in Wastewater Treatment Processes

Track S03. Modelling and Simulation in Applied Energy

Track S04. Modelling and Simulation in Solar Thermal Power Plants

Track S05. Object-Oriented Technologies of Computer Modelling and Simulation of Complex Dynamical Systems

Track S06. Chemical Process Systems Simulation

Track S07. Industrial Optimization Based on Big Data Technology and Soft Computing

Track S08. Simulation as Enabler for Innovative Technology

Track S09. Cooperative Automation

Panel discussions

Panel 1: Modelling and Simulation in Cleantech

Chairs:
Bernt Lie, University College of Southeast Norway, Norway
Jesús Zambrano, Mälardalen University, Sweden

Panelists:
Luis J. Yebra, CIEMAT-Plataforma Solar de Almería, Spain
Erik Dahlquist, Mälardalen University, Sweden

What is cleantech? Cleantech consists of products and services which are focused on the use of renewable natural resources and recycled materials in an energy-efficient way. Cleantech utilizes biological natural resources and turns them into food, energy, and other products and services. Cleantech uses clean technologies, which saves the environment by efficient recycling of materials. How? We have a broad range of technologies related to recycling, renewable energy, information technology, green transportation, electric motors, green chemistry, lighting, grey water, and more. Does this mean that cleantech is gradually introduced in all areas?

The environment is restored with pollution removal and avoidance. What can we do in practise? Air has been a focus area in industry, energy and traffic. Water treatment has been developed to remove undesirable chemicals, biological contaminants, suspended solids and gases from contaminated water. Where do we have the main risks? Availability of usable water may set constraints on operation. In industrial processes, closed water circulation is a goal which is beneficial for the environment. Wastewater treatment is needed for purifying contaminated water before returning it to the nature. Why are there difficulties in combining industrial and domestic wastewater treatment? Mining introduces many challenges for the environment. Renewable energy, including wind power, solar power, biomass, hydropower, biofuels etc., is an essential part in integrating cleantech with the energy production. Waste can be used as raw material or fuel in many ways. Power plants can use waste in energy production. What are the main challenges?

A circular economy aims to close the loop to make economy more sustainable and competitive. This should be more than just recycling. What does this mean? Water and wastewater treatment are good examples. There are challenging tasks for Information technology, modelling, control and optimisation. How can we proceed? What kind of Modelling and Simulation is important in cleantech? How can we compare alternative solutions and build situation awareness? The problem solving in cleantech includes the smart integration of all the historical elements, earth, fire, water and air, with data.

Panel 2: Future Energy Systems

Chairs:
Erik Dahlquist, Mälardalen University, Sweden,
Cristian Nichita. University of Le Havre, France

Panelists:
Rebei Bel Fdhila, ABB Corporate Research, Sweden
Luis J. Yebra, CIEMAT-Plataforma Solar de Almería, Spain Panelists:

A thermal power station or a coal fired thermal power plant is by far, the most conventional method of generating electric power with reasonably high efficiency. Technology has reached very high levels and environment is in focus in many ways. Bioenergy takes an increasing portion of the production: a wide variety of materials are used as fuels. Oil and gas hold a very strong position in overall energy usage. Biofuels provide new competing alternatives. CO_2 capture has taken a high role in research. Is it important also in practise? Are we going to bioeconomy? Is the thermal power a necessity in our energy balance?

Sustainable or renewable energy is considered as a future source of energy, but it is already strong in many forms: water power is well integrated in the energy system; solar and wind are getting more popular; geothermal, wave and tide energy can be locally very important. Electricity is increasingly popular both in solar and wind power. To what level it is sufficient? Efficiency is not very high in solar panels. Wind power cannot reach sufficient operating hours. We need storages but can we find practical solutions? Solar thermal power plants, especially concentrating technology, provide higher efficiency. There are many feasible solutions to thermal storage. What to use? How to design a system? What is needed in control? There are unavoidable disturbances.

Where do we use energy? Industry needs high reliable levels. Is the nuclear power a solution? Adaptation is easier in domestic use, but how to do it? Heating and cooling take the highest part. Solar energy can help but needs storage. Geothermal can be used as storage. What is the potential of buildings as storages? Do we need small scale Combined Heat and Power (CHP)? District heating systems are good solutions to bring the thermal energy to buildings. Smart grids have studied mainly for electricity. What do we need for smart thermal grids? In northern areas, we have consumption peaks. Can we cut them with smart adaptation? Traffic is under change: electricity is gaining popularity; interesting biofuels have been introduced; fuel cells are considered as a future option in the way to the hydrogen economy. How to integrate these with sustainable energy? How to choose an operable portfolio from the increasing alternatives of energy production?

Panel 3: Intelligent Systems and IoT in Future Automation

Chairs:
Esko K. Juuso, University of Oulu, Finland
Lars Eriksson, Linköping University, Sweden

Panelists:
Roy Calder, Schneider Electric, United Kingdom
Yukinori Suzuki, Muroran Institute of Technology, Japan
Galia Weidl, Daimler AG, Germany

In industry, intelligent systems have been developed for integrating data and expertise to develop smart adaptive applications. Recently, big data, cloud computing and data analysis has been presented as a solution for all kinds of problems. This provides feasible new things in global business and digitalisation in new applications. Can we take this as a general solution for automation? Are sensors only for collecting data to clouds? However, e.g. condition monitoring introduces huge volumes of data. Wireless solutions are improving fast: 3G, 4G, 5G. But can we transfer signals to clouds and store the data? Is this too much? Where is the expertise? Obviously, local calculations are needed. Are they based on intelligent systems? Also the security of the automation becomes increasingly important in distributed systems.

Transport systems are analysed as discrete event systems to find bottlenecks and avoid risks. Urban traffic is becoming an important area. Autonomous driving is a hot topic. What is needed to embed this in the urban traffic? Are there analogies with industrial systems? Mechatronics is an essential part in machines and many process devices. IoT with sensor development and access to traffic information opens up many opportunities for planning and control of transport through optimization.

What are the main differences between industrial systems and transport systems? Can we use similar control solutions? What can we learn from other areas? Can we find analogies? What is common? Where do we have differences? What kind of models do we need? What should the control

Author index

Modelling and Simulation of the Electric Arc Furnace Processes

Vito Logar

Faculty of electrical engineering, University of Ljubljana, Slovenia, vito.logar@fe.uni-lj.si

Abstract

Market demands on steel quality, price and production times dictate an introduction of technological innovations regarding the electric arc furnace (EAF) steelmaking. A developing field with significant potential is related to the advanced software support of the EAF operation, combining monitoring and proper control of the EAF. Such systems include process models, capable of continuous estimation of the unmeasured process values, such as chemical compositions and temperatures of the steel, slag and gas. The paper briefly presents the features of all developed EAF models, which are used together with the measured EAF data to estimate the unmeasured process values. The models are mainly implemented using non-linear, time-variant ordinary differential equations. The parameterization of the models was performed using the available EAF data, such as temperatures, steel and slag compositions, melting programs etc. The validation results that were performed using measured EAF data indicate high levels of estimation accuracy of all crucial steel-recycling values and processes. The accuracy of the presented models is in the range of +/- 15 K for steel temperature and +/-10 % for steel composition. Thus, accuracy of the models allows them to be used in broader software environments, such as soft sensors for process monitoring, optimization and operator decision support.

Keywords: electric arc furnace, EAF, modelling, simulation, validation

1 Introduction

Current market demands on steel quality, price and production times dictate an introduction of several technological innovations regarding the electric arc furnace (EAF) steelmaking. An emerging field with huge potential, but yet rather unexplored, is also advanced software support of the EAF operation. Running in parallel to the EAF process such systems allow online monitoring, fault detection and even model-based control of the process. Using such systems in parallel to the actual EAF process can have several advantages in comparison to the manual EAF operation, arising from the nature of the steel-melting process. As known, several crucial process values are hard to measure continuously, such as temperatures and chemical compositions of the steel, slag and gas etc. Using EAF process models, which integrate all significant EAF phenomena and use available EAF data to calculate the missing process values, results in a system, which is able to estimate the process values with sufficient accuracy. In this manner, better insight to the melting process can be established and consequently a more optimal operation of the EAF can be performed.

The paper presents an overview of the proposed EAF model, including electrical, hydraulic, mass-transfer, heat-transfer and chemical submodules in terms of modelling approach, modelling detail and schematic representations of the model structure. Since the mathematical models of the EAF have already been developed, validated and extensively described (Logar *et al* 2011, 2012) the paper presents only the key characteristics of each separate submodule and its importance for the overall accuracy of the calculations. Furthermore, the paper discusses possible and necessary upgrades of the models to implement them in process-optimization and decision-support frameworks. The aim of those is to present an EAF operation-support tool, which will be running in parallel to the EAF process and will be used for enhancement of the operation, such as: a) EAF operation monitoring based on soft-sensing technology, allowing a better insight to the melting process and consequently more optimal control of the EAF; b) process-optimization based on process models, optimizing the melting programs, according to the current state in the EAF; c) operator decision support, combining the advantages of model-based soft sensors and process optimization in one solution, representing the highest level of EAF software support.

2 Modelling of the EAF processes in general

Section The literature review in the field of modelling, optimization and control of the electric arc furnaces shows that many different models and engineering approaches studying the EAF processes have been developed. Most of these are focused on particular processes of the EAF and were developed for the purpose of the field research or simulation of the EAF operation. A few models were designed especially for their implementation into industrial applications as an operator-support tool for monitoring of the recycling process and thus easier decision making and control of

the processes. First models associated with the EAF processes were introduced back in 1980s and were functionally extremely limited. Modern models have progressed in their complexity, usability and accuracy and are also used in industrial applications for monitoring of the EAF during the steel-melting process. Below, some of the most relevant research and practical applications in the field of modelling, optimization and control of the EAF processes are described.

Woodside (1970) introduces a concept for optimal EAF control, based on Pontryagin approach and uses it for optimization of the energy input during coke injection. The simplified model was introduced in 1970 and was able to estimate bath temperature and carbon concentration. Montanari *et al* (1994), Tseng *et al* (1997), Collantes-Bellido and Gomez (1997) introduce mathematical models describing the electrical part of the EAF and the impact of the EAF operation on electrical grids in terms of disturbances (flickers) and their elimination.

In 1999, Bekker *et al* (1999) develops a mathematical model implementing thermodynamic relations for the purposes of EAF-process simulation. The model is simplified and assumes that all available heat is transferred directly to the steel bath and further from the bath to the solid steel. Although it addresses some important chemical reactions and the released energy, the presented simulation results are not validated and thus applicable only with limitations. Nonetheless, the Bekker model represents one of the first attempts to model all crucial processes of the EAF. Additionally, Bekker *et al* (2000) introduces a concept of model-predictive EAF control (MPC).

Oosthuizen *et al* (2001, 2004) designs a mathematical model of the EAF processes derived from the Bekker model. Using a more complex modelling approach, the proposed model gains on estimated offgas temperature accuracy and allows a calculation of the slag foaming. Furthermore, optimal controller is introduced, which should control the furnace in a manner to reduce its operational costs. Similarly to Bekker, a simulation study involving a model-based control (MPC) is performed on the model for the purposes of cost reduction.

One of the most sophisticated EAF models up to 2005 was introduced by MacRosty and Swartz (2005, 2007) and used for optimization of the EAF process. The model considers the complete EAF and includes chemical, mass- and heat-transfer processes. Due to modelling simplifications, the EAF layout is divided into four zones with similar physical characteristics. Chemical reactions in each zone are based on molar-mass equilibria, while the overall model is based on energy and mass equilibrium. The model was validated using the measured operational data and can be used to estimate bath temperature, bath composition and slag composition. Further on, the model is implemented in a simulation study to optimize the operational costs of the EAF. The authors report of several issues regarding the optimization procedure and its unreliability.

Logar *et al* (2011, 2012) introduce complex EAF models, including electrical, hydraulic, chemical, heat- and mass- transfer processes. The models are based on fundamental physical laws and are validated on measured operational data of the EAF. The results show high levels of similarity between the measured and the simulated data. The combination of all developed models in one functional model represents the most complex approach to EAF modelling found in literature. Many studies have been performed in the field of numerical modelling of the EAF. The field has been established as a new, fast emerging science and engineering discipline that encompasses computational solid mechanics (Fung *et al*, 2001) and fluid mechanics (Schiestel, 2008), connected with solidification phenomena (Dantzig et al, 2009, Glicksman, 2010), allotropic transformations phenomena (Hazotte, 2003), and put into the context of computational microstructure evolution (Janssens *et al*, 2007) in processes like casting (Fredriksson and Åkerlind, 2006), heat treatment (Gür and Pan, 2009) etc. The use of numerical modelling proved to be an efficient approach for modelling the relations between bath stirring, fluid flow and electromagnetic (EM) forces. Due to the complex coupling between flow and EM forces, numerical modelling is the most economical way of analyzing, optimizing and developing new applications. Many modern industrial processes, such as electrical arc furnaces, rely on findings of EM, heat-, mass-transfer and metallurgical science. Their interconnections are currently not sufficiently understood and computationally modelled.

Furthermore, modelling approaches to gas-phase phenomena (Meier *et al*, 2015, Kolagar *et al*, 2015) and EAF off-gas heat recovery have been proposed (Gandt *et al*, 2016).

The implementation of the mathematical models in industrial applications can be found at renowned EAF manufacturers and users, such as Tenova, Siemens VAI, SMS Siemag, Centre de Recherche Métallurgiques (CRM), ArcelorMittal and BFI (Clerici *et al*, 2008, Dorndorf *et al*, 2007, Khan *et al*, 2013, Natschläger *et al*, 2008 and Nyssen *et al*, 2007). The developed models vary in modelling approach, modelling detail, usability, accuracy and types of input data used for estimating the process values. Reviewing the available literature, most of these models are based on static calculations using statistical or regression methods, i.e. SMS Siemag, BFI and Tenova. The model, based on dynamic modelling approach was introduced by CRM. It relies on fundamental laws of thermodynamics and is used to estimate the end-point bath temperature. It is claimed that the accuracy of the model is +/- 50 K, which

indicates that further improvements of the model are possible.

The literature review reveals that all models used in industrial applications are designed solely for estimation of the process values, while no support to the operators in the sense of optimal control is given. Thus, the main challenge in EAF operation, i.e. determination of optimal melting programs, times and amounts of charged materials, is therefore still left to the operator and his experience.

Regarding the literature review, a design of the overall EAF process model was initiated, incorporating all crucial EAF processes and focusing on accuracy and usability of the obtained solution for further development and its inclusion to several other software environments.

3 Modelling approach

3.1 General

The models as presented in this paper have been developed according to the fundamental physical laws by means of non-linear, time-variant, first order differential equations; although, several other approaches could be implemented as well. The selected approach has its advantages and drawbacks when compared to other possibilities; however, the possibility to use the developed models with as many EAF designs as possible was the main aim of the study and for this reason the models are based on fundamental mathematical/physical approaches. The model can be presented schematically as in Fig. 1.

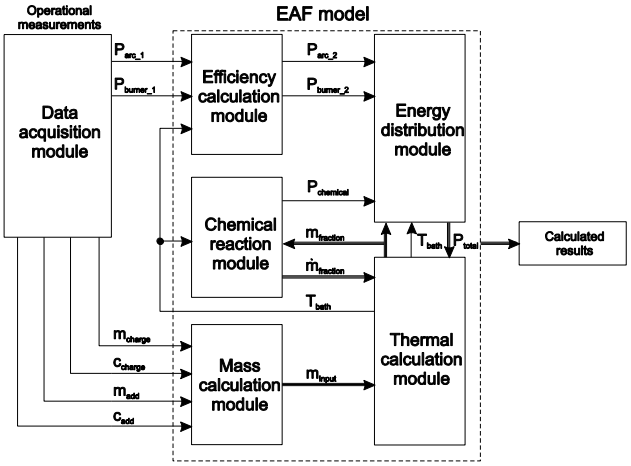


Figure 1: Schematic presentation of the developed EAF models

The presented model as shown in Fig. 1 comprises mathematical descriptions of all main physical processes appearing during the steel-recycling process, i.e. electrical, hydraulic, thermal (including radiation), chemical and mass transfer. As shown in Fig. 1, the model for estimation of the EAF process values is composed of several modules, which contain mathematical relations describing the physical properties of the EAF steel melting process and the corresponding model parameters. The calculations are grouped in submodules in order to simplify the model structure and assure low computational loads.

Due to complexity of the modelled processes and in order to simplify the obtained models, the EAF layout has been divided into several zones, assuming that each zone is homogenous and possesses equal physical characteristics, such as temperatures, densities, heat transfer coefficients etc. The zones used in the model are solid steel, liquid steel, solid slag, liquid slag, gas, roof and walls, as shown in Fig. 2.

According to the above, calculations of a separate submodule are limited only to certain zones, i.e. electrical and hydraulic models appear in no zone directly, heat-transfer model appears in all zones, mass transfer model appears in solid steel, liquid steel, solid slag, liquid slag and gas zones, and the chemical model appears in liquid steel, liquid slag and gas zones.

3.2 Model characteristics

Each of the above models utilizes different physical laws and mathematical equations in order to obtain the values needed by other models or as an end result/estimation. The electrical and hydraulic models can be characterized by the following:

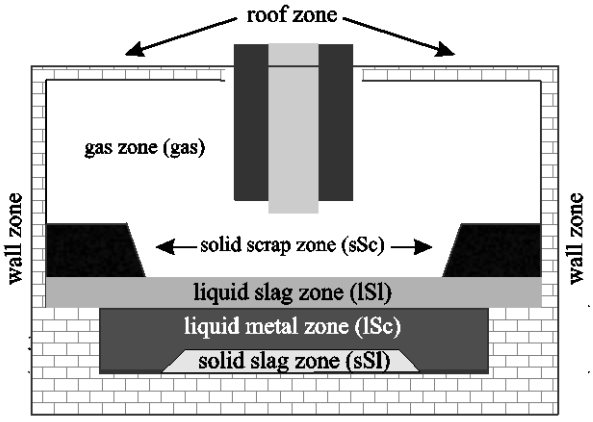


Figure 2: Division of the EAF layout to different zones

- all electrical values are calculated using harmonic analysis, i.e. in complex space,
- Cassie-Mayr arc model (1st order ODEs) is used with additional variable Lorentz noise,
- the models utilize transformer and reactor taps, resistances/reactances of lines, transformer, arcs and steel, all electrical values (voltages, currents, powers, energies, power factors, impedances etc.),
- electrode control is carried out using a hydraulic model and three independent PI controllers,
- the model parameters are variable for different stages of the melting process, i.e. electrode bore-down, melting, flat bath.

The heat-transfer model can be characterized by the following:

• the melting process is divided to different phases of the melt-down (electrode bore-down, exposing panels, flat bath etc.)

• 1st order ODEs are used to calculate the temperatures based on energy input/output balances

• heat-transfers are utilized to each zone from: arcs, burners, chemical reactions, volatile oxidation, electrode oxidation and other zones,

• heat losses are utilized due to cooling of the furnace, offgas extraction, steel and slag enthalpy,

• implementation of geometry supported (view-factor based) radiative heat exchange,

• taking into the account temperature-dependent burner efficiency and continuous transitions between the zones (geometry supported).

The mass-transfer model can be characterized by the following:

• the melting process is divided to different phases of melt-down (electrode bore-down, exposing panels, flat bath etc.)

• 1st order ODEs are used to calculate mass transfers based on temperature levels (melting) and energy input/output balances,

• elements and compounds which are taken into the account in calculations are:

- steel zone: Fe, C, Si, Cr, Mn, P

- slag zone: FeO, SiO_2, MnO, Cr_2O_3, CaO, MgO, Al_2O_3, P_2O_5

- gas zone: N_2, O_2, CO, CO_2, CH_4 (gas burners),

• implementation of reversible dynamics (cooling and solidification of the steel),

•calculation of mass transfers due to: melting, charging and slag addition, oxygen-fuel burners, oxygen lancing, carbon injection and chemical reactions.

The chemical model can be characterized by the following:

• implementation of all main chemical reactions appearing in the steel-melting process (oxidation/reduction of Fe, FeO, Si, SiO_2, C, CO, Mn, MnO, Cr, Cr_2O_3, P and P_2O_5),

• 1st order ODEs are used to calculate rates of change of elements/compounds based on molar equilibria with reaction equilibria constants dependent on molar composition of the zone,

• utilization of chemical energy exchange due to exothermic and endothermic reactions,

• calculation of foaming slag height based on slag density/viscosity/surface tension and superficial gas velocity (CO) including slag decay,

• calculation of online and endpoint steel/slag/gas compositions and relative pressure.

The parameters of the model (approximately 100) were obtained using known data or conclusions of different studies (transformer taps, resistances/reactances, furnace dimensions, heat capacity coefficients, densities, emissivities, enthalpies of formation, reaction rates, molar masses etc.) or were determined experimentally using the available initial, online or endpoint operational EAF measurements (cathode voltage drops, arc temperatures, arc conductances, arc cooling constants, slag-reactance coefficients, heat-transfer coefficients, specific area coefficients, arc-energy distributions etc.). The validation of the models was carried out using operational EAF measurements, which were obtained during different melting scenarios. In this manner, an accurate model of the actual EAF recycling process was obtained.

4 Results

The displayed results show the most important estimations of the process values while operating the EAF, i.e. bath temperatures, steel compositions and slag compositions. Fig. 3 shows the comparison between measured and model simulated values for initial steel mass, endpoint steel mass, power on time and bath temperatures. The results were obtained from several heats and are represented in a form of a mean value with standard deviation.

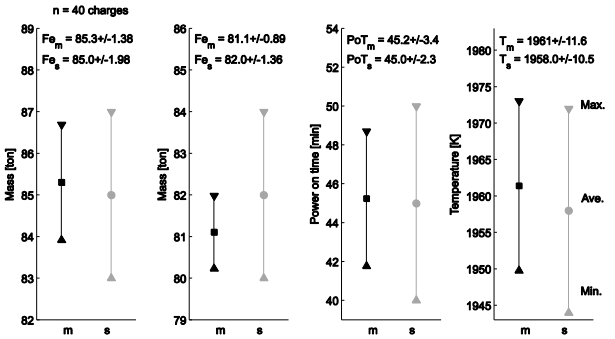


Figure 3: comparison between measured and simulated values for initial (1st) and enpoint (2nd) steel mass, power on time (3rd) and bath temperatures (4th). Black squares and grey circles represent measured and simulated mean values, while black and grey triangles represent measured and simulated standard deviations, respectively.

As can be seen in Fig. 3, all measured and simulated values are similar, both in mean values and in standard deviations. The most important validation values from Fig. 3 are steel yield (difference between the initial and endpoint steel masses) and steel bath temperature. Bath temperature is usually measured one to three times before tapping, while steel yield is determined at

tapping. Neither of these values is measured continuously during the EAF operation.

Fig. 4 shows the comparison between measured and model simulated endpoint steel composition in a form of a mean value with standard deviation.

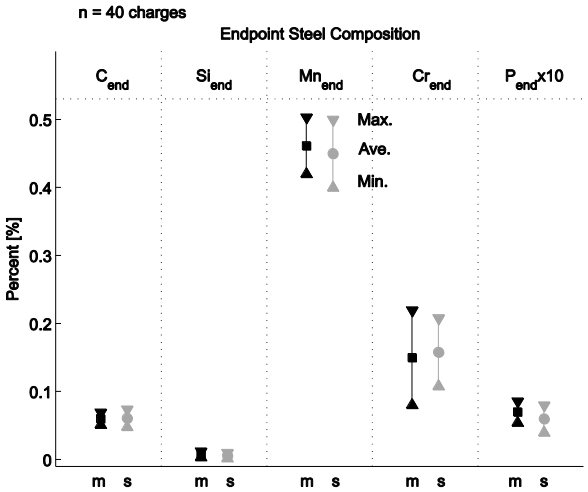


Figure 4: comparison between measured and simulated endpoint chemical compositions of the steel

As can be seen in Fig. 4, all measured and simulated values are similar, both in mean values and in standard deviations. The most important validation value from Fig. 4 is the carbon content in the steel, since carbon percentage is (among others) directly linked to different steel grades produced and has to be determined and contained in proper amount. Complete steel composition is determined at tapping and is otherwise not measured continuously.

Fig. 5 shows the comparison between measured and model simulated endpoint slag composition in a form of a mean value with standard deviation.

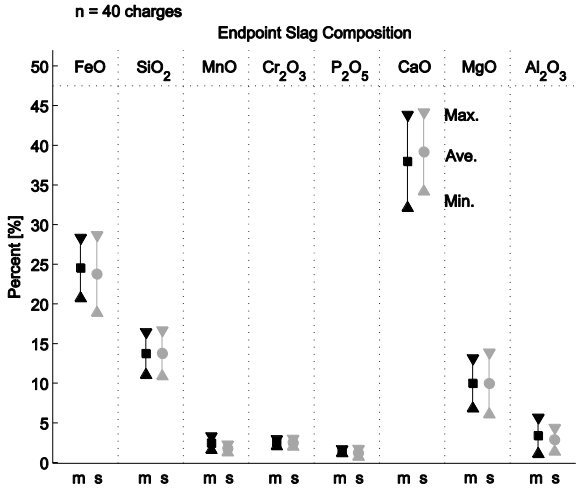


Figure 5: comparison between measured and simulated endpoint chemical compositions of the slag

As can be seen in Fig. 5, all measured and simulated values are similar, both in mean values and in standard deviations. The most important validation value from Fig. 5 are FeO, SiO2, MnO, CaO and MgO contents in the slag, since these compounds define the properties of the slag, which are linked to its foaminess and protective characteristics.

Fig. 6 shows the energy balance as calculated by the proposed model.

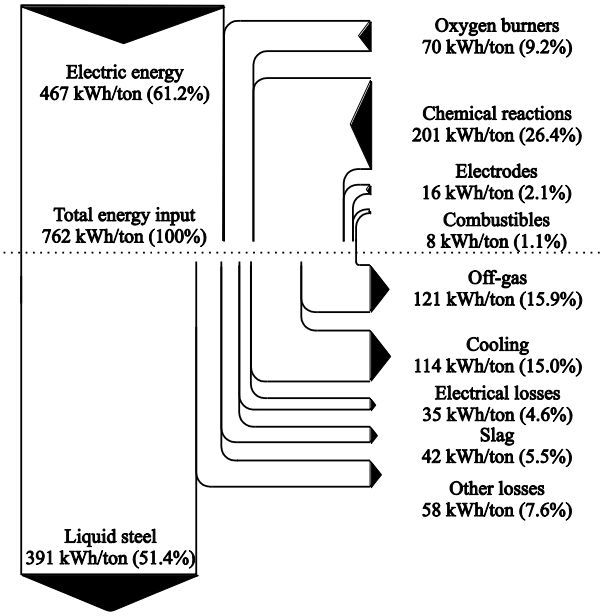


Figure 6: energy balance of the EAF as obtained by the proposed model

As can be seen in Fig. 6, total energy input required to produce 1 ton of steel is approximately 760 kWh. More than a half of this energy is represented by electrical energy, while other important sources of energy are oxygen burners and chemical reactions. Regarding the losses of energy, approximately 390 kWh of energy is held by steel enthalpy due to its high temperature. This energy is later lost to the environment as the steel cools down. Furthermore, off-gas extraction and cooling of the furnace vessel also represent an important sinks of energy.

5 Practical applications of the model

Due to the nature of the EAF steel-recycling process (high temperatures and electric currents), performance of the crucial process measurements is difficult. Consequently, monitoring and control of the melting process is performed using the operator's experience and is based on indirect measurements (e.g. power-on time, consumed energy, arc stability etc.) and not on the actual conditions in the EAF (e.g. stage of melting, bath composition, bath temperature), which leads to suboptimal operation, i.e. lower energy and raw material efficiency, increased off gas and CO2 emissions, decreased quality of the steel; and consequently higher operational costs.

Furthermore, operational efficiency is influenced also by variable composition of the input materials (steel scrap, non-metallic additives). The fluctuations in EAF operation can be resolved using a combination of EAF

process models, optimization techniques and decision support methods. The combination of these methods together with available process measurements forms a supporting system for operation of the EAF. Such a system uses process measurements as inputs, in order to provide a better insight into the current EAF conditions and to suggest the most appropriate action, leading to more efficient operation of the EAF. Using mathematical models, which are designed in compliance with the physical laws and using available measurements as inputs, crucial process values, which are not measured, can be estimated in parallel to the EAF process with high accuracy.

In this manner, an optimal operation of the EAF can be established, leading to higher steel yield, lower energy, raw material and additive consumption, shorter production times, higher steel quality etc. The introduction of such operation indirectly leads to improved economic, ecological and technological aspects of the mills, with such system installed.

6 Conclusions

In this paper a brief explanation of the modelling approach to crucial EAF processes as well as its potential use in higher-level applications is presented. Furthermore, some key comparisons between the measured and the simulated values are presented, showing the overall accuracy of the calculations.

The objective of developing a complete EAF model is to use it in application frameworks for different purposes, such as online monitoring, process optimization or operator decision support. Since the description and modelling details are far too great and extend the frame of this paper, all interested readers can refer to the reference list (Logar et al.) for further information.

References

J. G. Bekker, I. K. Craig, P. C. Pistorius. Modeling and simulation of an electric arc furnace process, *ISIJ International*, 39(1): 23 – 32, 1999.

J. G. Bekker, I. K. Craig, P. C. Pistorius. Model Predictive Control of an Electric Arc Furnace Off-Gas Process, *Control Engineering Practice*, 8(4): 445 - 455, 2000.

P. Clerici, F. Dell'Acqua, J. Maiolo, V. Scipolo. Dynamic EAF control using highest performing technology, *MPT International - Metallurgical Plant and Technology*, 31: 40 - 42, 2008.

R. Collantes-Bellido, T. Gómez. Identification and modelling of a three phase arc furnace for voltage disturbance simulation. *IEEE Transactions on Power Delivery*, 12(4): 1812 - 1817, 1997.

J. Dantzig, M. Rappaz. *Solidification*, CRC Press, Boca Raton, 2009.

M. Dorndorf, W. Wichert, M. Schubert, J. Kempken, K. Krüger. Ganzheitliche Erfassung und Regelung des Schmelzprozesses eines Elektrolichtbogenofens, *Stahl und Eisen*, 127: 63 – 71, 2007.

H. Fredriksson, U. Åkerlind. *Materials Processing During Casting*, John Wiley, Hoboken, 2006.

Y.C. Fung, P. Tong. *Classical and Computational Solid Mechanics*, World Scientific, Singapore, 2001.

K. Gandt, T. Meier, T. Echterhof, H. Pfeifer. Heat recovery from EAF off-gas for steam generation: analytical exergy study of a sample EAF batch, *Ironmaking & Steelmaking: Processes, Products and Applications*, 43(8): 581 – 587, 2016.

M. E. Glicksman. *Principles of Solidification*, Springer Verlag, Berlin, 2010.

C. H. Gür, J. Pan (eds.). *Thermal Process Modelling of Steel*, CRC Press, Boca Raton, 2009

A. Hazotte (ed.). *Solid State Transformation and Heat Treatment*, Wiley-VCH, Weinheim, 2003.

K. G. F. Janssens, D. Raabe, E. Kozeschnik, M. A. Miodownik, B. Nestler. *Computational Materials Engineering - An Introduction to Microstructure Evolution*, Elsevier, Amsterdam, 2007.

M. Khan, S. Mistry, V. Scipolo, S. Sun, S. Waterfall. Next-generation EAF optimization at ArcelorMittal Dofasco Inc, In *AISTech, Iron and Steel Technology Conference and Exhibition*, pages 697 – 706, Pittsburgh, USA, 2013, AISTech.

A. H. Kolagar, T. Meier, T. Echterhof, H. Pfeifer. Application of genetic algorithm to improve an electric arc furnace freeboard model based on practical data, International *Journal of Engineering Systems Modelling and Simulation*, 7(4): 244 – 255, 2015.

V. Logar, D. Dovzan and I. Škrjanc. Mathematical modeling and experimental validation of an electric arc furnace, *ISIJ International*, 51(3): 382 – 391, 2011.

V. Logar, D. Dovzan and I. Škrjanc. Modeling and validation of an electric arc furnace: Part 1, heat and mass transfer, *ISIJ International*, 52(3): 402 - 413, 2012.

V. Logar, D. Dovzan and I. Škrjanc. Modeling and validation of an electric arc furnace: Part 2, thermochemistry, *ISIJ International*, 52(3): 414 - 424, 2012.

V. Logar and I. Škrjanc. Modeling and validation of the radiative heat transfer in an electric arc furnace, *ISIJ International*, 52(7): 1225 - 1232, 2012.

R. D. M. MacRosty, C. L. E. Swartz. Dynamic modeling of an industrial electric arc furnace, *Industrial & engineering chemical research*, 44(21): 8067 – 8083, 2005.

R. D. M. MacRosty, C. L. E. Swartz. Dynamic optimization of electric arc furnace operation, *AIChE Journal*, 53(3): 640 – 653, 2007.

T. Meier, A. H. Kolagar, T. Echterhof, H. Pfeifer. Gas phase modeling and simulation in an electric arc furnace process model for detailed off-gas calculations in the dedusting system, In *Steelsim - International Conference Simulation and Modelling of Metallurgical Processes in Steelmaking 2015*, Bardolino, Italy, 2015, AIM.

G. C. Montanari, M. Loggini, A. Cavallini, L. Pitti, D. Zaninelli. Arc-furnace model for the study of flicker compensation in electrical networks, *IEEE Transactions on Power Delivery*, 9(4): 2026 - 2036, 1994.

S. Natschläger, S. Dimitrov, K. Stohl. EAF process optimization: theory and real results, *Archives of Metallurgy and Materials*, 53(2): 373 – 378, 2008.

P. Nyssen, G. Monfort, J. L. Junque, M. Brimmeyer, P. Hubsch, J. C. Baumert. Use of a dynamic metallurgical model for the on-line control and optimization of the electric arc furnace, In *Steelsim - International Conference Simulation and Modelling of Metallurgical Processes in Steelmaking 2011*, Graz, Austria, 33-38, 2011, AIM.

D. J. Oosthuizen, J. H. Viljoen, I. K. Craig, P. C. Pistorius. Modeling of the off-gas exit temperature and slag foam depth of an electric arc furnace, *ISIJ International*, 41(4): 399 - 401, 2001.

D. J. Oosthuizen, I. K. Craig, P. C. Pistorius. Economic evaluation and design of an electric arc furnace controller based on economic objectives, *Control engineering practice*, 12(3): 253 - 265, 2004.

R. Schiestel. *Modeling and Simulation of Turbulent Flows*, Willey, Hoboken, 2008.

K. J. Tseng, Y. Wang, D. M. Vilathgamuwa. An experimentally verified hybrid Cassie-Mayr electric arc model for power electronics simulations. *IEEE Transactions on Power Electronics*, 12(3): 429 - 436, 1997.

C. M. Woodside. Singular arcs occurring in optimal electric steel refining, *IEEE transactions on automatic control*, 15(5): 549 - 556, 1970.

Situation Awareness and Early Recognition of Traffic Maneuvers

Galia Weidl[1] Anders L. Madsen[2,4] Viacheslav Tereshchenko[1,3]
Wei Zhang[1,3] Stevens Wang[1,3] Dietmar Kasper[1]

[1]Department of Driving Automation, Daimler AG, Group Research & AE, 71034 Böblingen, Germany
{galia.weidl, dietmar.kasper}@daimler.com
[2]HUGIN EXPERT A/S, [4]Department of Computer Science, Aalborg University, Denmark
anders@hugin.com
[3]University of Stuttgart, Germany
{viacheslav.tereshchenko, zw158535165, stevens.rx.wang}@gmail.com

Abstract

We outline the challenges of situation awareness with early and accurate recognition of traffic maneuvers and how to assess them. This includes also an overview of the available data and derived situation features, handling of data uncertainties, modelling and the approach for maneuver recognition. An efficient and effective solution, meeting the automotive requirements, is successfully deployed and tested on a prototype car. Test driving results show that earlier recognition of intended maneuver is feasible on average 1 second (and up to 6.72 s) before the actual lane marking crossing. The even earlier maneuver recognition is dependent on the earlier recognition of surrounding vehicles.

Keywords: Bayesian networks, massive data streams

1 Introduction

A highway, typically providing several traffic lanes, is characterized by complex scenes with many vehicles. Reliable situation assessment requires multi-sensor fusion and management of uncertainty in order to interpret accurately the traffic environment. To reduce the risk of accidents and congestions, an autonomous system must analyze and be aware of possible hazards of a driving situation. This includes: correctly recognizing intended maneuvers of all surrounding vehicles at an early stage and using this information to enable corrective actions like braking or steering, thus helping to avoid or mitigate potential collisions. Situational awareness and recognition of traffic maneuvers are key elements of modern driver assistance and autonomous driving systems (Kasper *et al*, 2011, 2012, 2013; Morris *et al*, 2011; Kumar *et al*, 2013; Tereshchenko, 2014; Schlechtriemen *et al*, 2014; Weidl *et al*, 2014, 2015, 2017; Mori *et al*, 2015; Satzoda *et al*, 2015; Zeisler *et al*, 2015).

A probabilistic approach, using Object-Oriented Bayesian networks (OOBNs) for maneuver recognition has been proposed in (Kaper *et al*, 2011, 2012, 2013) and (Zeisler *et al*, 2015). It is based on the own (ego) vehicle dynamics, its driving path in relations to the lane markings and/or surrounding vehicles, to evaluate the vehicles' relevance as possible target objects and to recognize earlier maneuvers in real traffic. In addition to the data from in-vehicle sensors, including both the vehicle kinematics and vehicle surround dynamics, (Satzoda *et al*, 2015) also uses visual data from multiple perspectives to characterize lane changes. In (Kaper *et al*, 2011, 2012, 2013), we use pairwise vehicle-vehicle relations; as far as the sensors can percept the surrounding objects.

In (Mori *et al*, 2015) an approach, based on Hidden Markov Models computes the driver's intent on lane change and/or potential risk of accidents. This work studied the driver decisions whether is safer to change lane in front of a faster car closing the gap or to brake for keeping a safe distance to a slower car in front, by evaluating the time-to-collision (TTC). We have used similar features in the safety model (in section 3) to infer the intention on lane change for the situations shown in Figure 1, utilizing the relative longitudinal dynamic between a vehicle (own, neighbor) and the vehicles in front and back, driving on the target lane (Tereshchenko, 2014). This effectively builds a gap for finishing a lane change maneuver and has an impact on the decision for lane changing, see (Tereshchenko, 2014; Mori *et al*, 2015; Weidl *et al*, 2015; Yan *et al*, 2015). The mentioned features, characterizing the vehicle state, were extended in (Li *et al*, 2015) with the driver's operation signals to enhance by hidden Markov models the classification of lane change maneuvers.

In our recent work (Weidl *et al*, 2015) and for the results reported here, we have focused on the development of a solution for maneuver recognition and its deployment on a Linux based system emulating the automotive target platform, using commercially viable sensors, image processing and multi-sensor fusion. We use the commercial software HUGIN, allowing efficient BN modelling and automated c-code generation. All

current and future developments are compared to our initially developed "Original (ORIG)" OOBN, which has shown promising results as described in (Kasper *et al*, 2011, 2012, 2013; Weidl *et al*, 2014). Our latest work (Weidl *et al*, 2015) describes three statistical classifiers as deployed on the prototype vehicle as well as the planned dynamic extension into a Dynamic Bayesian Network (DBN). The DBN are now deployed on the vehicle and evaluated in highway drive and in statistical comparison to other static classifiers. The main contribution of this work is the extension of all four classifiers with special evidence for better accuracy. The DBN deployment on the automotive target platform has been successfully tested in real highway driving.

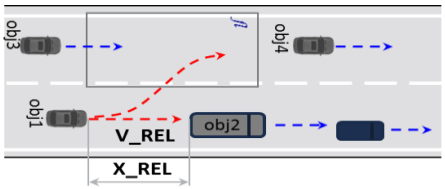


Figure 1. *Longitudinal* relative dynamics between following and front vehicles, both moving on the same lane. *Lateral* relative dynamics towards the lane marking, when initiating a lane change maneuver.

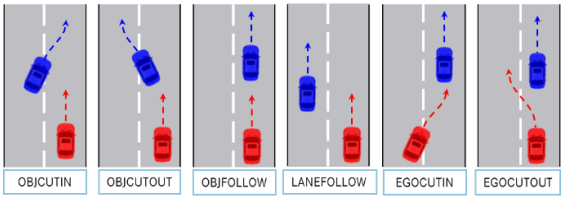


Figure 2. Maneuver recognition as a vehicle-vehicle relation between the own EGO-vehicle (red) and a neighbor OBJ-vehicle (blue).

The automotive target platform represents both a storage and an inference challenge. The requirement on automotive safety demands accuracy close to 100% for a prediction horizon of 1 second. Moreover, the solution should scale to the specification of the hardware restrictions of the target platform. It should meet the automotive requirements on computation time and memory space, which are strongly constrained by the electronic control unit. The quick development of situations over time requires an automatic system, capturing and analyzing massive data streams, under uncertainties, every 20 milliseconds for several safety applications, resulting in 0.15 milliseconds for maneuver recognition.

This paper is organized as follows. Section 2 gives an overview of the used method, section 3 – on efficient modeling, section 4 - on proper treatment of data uncertainties and their use for building of hypotheses on driving behavior for event (maneuver) recognition. The approach is outlined in section 5, while section 6 describes the evaluation of classifiers. Section 7 summarizes the results with outlook.

2 Method for Probabilistic Reasoning

2.1 Bayesian network (BN)

A Bayesian network $\boldsymbol{BN:= (G, P)}$ is defined as a directed acyclic graph $\boldsymbol{G}$ and $\boldsymbol{P}$ - a set of CPDs (conditional probability distributions) $\boldsymbol{P(X \mid pa(X))}$ of a variable X conditioned on its influence variables $\boldsymbol{pa(X)}$, (Friedman and Koller 2009, Kjærulff and Madsen 2013). The joint probability of a BN is computed by the *Chain rule for BNs*:

$$\boldsymbol{P(X_1, \ldots, X_n) = \Pi_{i=1..n}\, P(X_i \mid pa(X_i))} \qquad (1)$$

The graph $G=(V,L)$ contains nodes V (to represent random variables) and edges L to represent the conditional dependency relations between the nodes. A BN can be used as a knowledge representation to compute the probability $\boldsymbol{P(X=x|e)}$ given a set of observations e. The Bayesian theorem allows inverting the probability computations, i.e.

$$P(X|Y) = P(Y|X)P(X) / P(Y) \qquad (2)$$

An *object-oriented Bayesian network (OOBN)* contains instance nodes in addition to the usual BN nodes. An instance node is an abstraction of a net fragment into a single unit (network class) (Friedman and Koller 2009, Kjærulff and Madsen 2013). Therefore, instance nodes can be used to represent different network classes as well as repetitive structures within other nets (encapsulation). Thus an OOBN can be viewed as a hierarchical (data/information fusion) model of a problem domain. Every layer in this hierarchy expresses another level of abstraction in the OOBN model. The modeling extensions in this work explore also *dynamic Bayesian networks (DBN)*, which use a time series of observations for information fusion and inference (Friedman and Koller 2009; Zhang and Ji 2009; Kjærulff and Madsen 2013; Weidl *et al*, 2015-2017). In this work, we use OOBN and DBN to represent the extension for both the lateral and longitudinal relative dynamics. DBN combine repetitive BN structures as discrete time slices. They follow the 1st order Markov assumption, i.e., the future $\boldsymbol{X^{t+1}}$ is independent on the past $\boldsymbol{X^{t-1}}$ given the present $\boldsymbol{X^t}$: $\boldsymbol{(X^{t+1} \perp X^{t-1} \mid X^t)}$ together with the stationary assumption, that the transitional probability distributions do not change between the time slices:

$$\boldsymbol{P(X^{t+1} \mid X^t) = P(X^t \mid X^{t-1})}.$$

3 OOBN for Maneuver Recognition

To recognize the maneuvers considering the relative vehicle-vehicle motion (Figure 2), we have modeled them as states of variable MNVR(≡Maneuver of Pairs)

at the top layer (Figure 3, Figure 4) of the OOBN. The pairwise combination of vehicles' maneuvers ensures scalability of the approach. It reduces the memory requirements and uses computation resources only for the actually present surrounding vehicles. The OOBN fuses in the hypotheses at the lowest level of abstraction (see Figure 3) features under uncertainties (Figure 5, Figure 6), i.e. measured multi-sensor data and computed situation features.

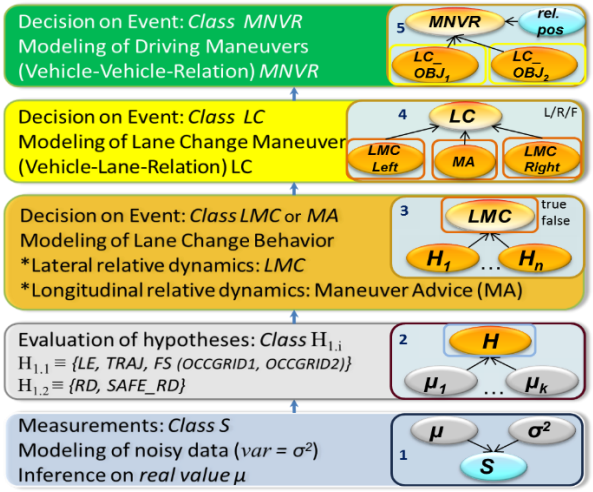


Figure 3. OOBN structure and layers for maneuver recognition.

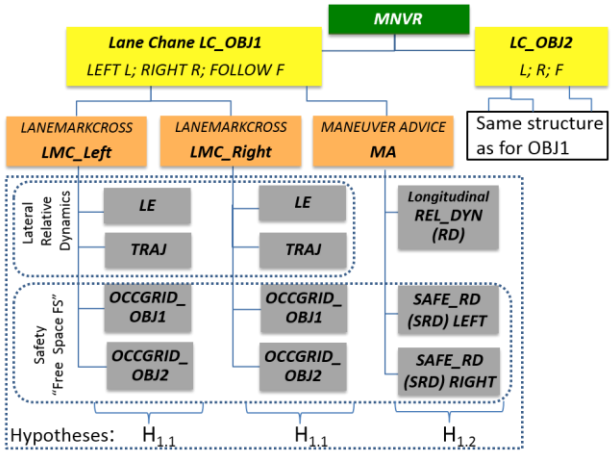


Figure 4. Class hierarchy of the OOBN for maneuver recognition between two vehicles (OBJ1 and OBJ2), expressed as lateral and longitudinal relative dynamics, including their safety (see Figure 1**).**

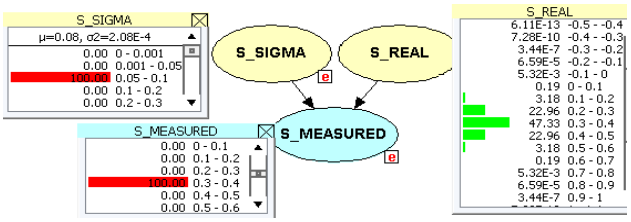


Figure 5. BN fragment for modeling of sensor's uncertainties with a discrete MEASURED variable.

4 Modeling of Data and Driving Behavior

4.1 Modeling of sensor data and its uncertainty

The maneuver recognition represents a task of the type reasoning under uncertainties with heterogeneous data. These data are acquired from multi-sensors measurements as well as from thereof fused and computed (by physical models) situation features. All data have naturally inherited uncertainties. The data characterize a traffic situation and define the set of situation features for maneuver recognition. The data input is represented as variables in the OOBN to support the inference process by allowing the measured (or computed) values of the variables to be inserted as evidence. For discrete variables, to be able to distinguish between the states of deduced features and to deal with the uncertainties in the sensor signals (Kaper *et al*, 2011, 2012, 2013), the measured signals are discretized in predefined partitions (Figure 5). In general, the measured signal $S_{measured} \equiv S_m$ is composed of its real (expected) value S_REAL $\equiv S_{expected}$ under measurement and its disturbance (sensor noise) S_{err} around the real value, i.e. $S_m = S_{expected} + S_{err}$. In many practical applications (and in our work), the sensor noise is assumed as a zero-mean Gaussian random process. Then, the disturbance is described by the signal variance $S_{err} \equiv S_\sigma^2$.

If the measurement instrument is not functioning properly (due to senor noise or fault), then the sensor-reading (S_MEASURED $\equiv S_m$) and the real variable (S_REAL) under measurement do not need to be the same. This fact imposes the causal model structure as shown in Figure 5, taking care of the uncertainties in the input data. The sensor-reading S_m of any measured variable is conditionally dependent on random changes in two variables: real value under measurement (S_REAL$\equiv S_\mu$) and sensor fault (S_SIGMA$\equiv S_\sigma^2$):

$$P(S_m \mid S_\mu, S_\sigma^2) = N(S_\mu, S_\sigma^2) \qquad (3)$$

where $N(S_\mu, S_\sigma^2)$ denotes the Gaussian distribution with mean value S_REAL $\equiv S_\mu$. Then, in principle, the probability distribution of the real value S_μ of the measured variable is inferred by equations (2) - (3), given the observation (evidence) from its sensor measurement S_m and its sensor disturbance S_σ^2. The last is obtained from the sensor diagnostics by use of a Kalman filter.

In the discrete case, the CPD of S_m, expressed as (3), is represented by a conditional probability table (CPT), while in the continuous case S_m is modeled by a continuous random variable with a linear continuous Gaussian (CG) conditional distribution function $N(S_\mu, S_\sigma^2)$. A BN with CG nodes is referred to as a Conditional

Linear Gaussian BN. It induces a multivariate normal mixture density of the form:

$$P(\Delta)\, f(\Gamma) = \Pi_{X \in \Delta}\, P(X \mid pa(X))\; \Pi_{X \in \Gamma}\, f(X \mid pa(X)),$$

where Δ are the discrete and Γ are the continuous variables.

The degree of uncertainty for a variable, which is *computed* from noisy sensor data measurements, is obtained as error estimation by variance calculus. For example, the uncertainty in velocity Δv and in distance measurement Δs directly affects the uncertainty of the computed time to reach a certain point at distance s with velocity v. The time is computed as a function of these two variables, i.e. $t = \frac{s}{v} = f(s, v)$. Its uncertainty value or its variance $t_\sigma^2 \equiv \delta t$ is computed for each vehicle object by taking the partial derivative of the time function $f(s, v)$:

$$\delta t = \sqrt{\left(\frac{\delta f}{\delta s}\right)^2 + \left(\frac{\delta f}{\delta v}\right)^2} = \sqrt{\left(\frac{\partial f}{\partial s}\delta s\right)^2 + \left(\frac{\partial f}{\partial v}\delta v\right)^2}$$
$$= \sqrt{\left(\frac{1}{v} \cdot \Delta s\right)^2 + \left(-\frac{s}{v^2} \cdot \Delta v\right)^2}$$

To model the uncertainty in a variable, which has been computed from noisy sensor measurements, we use Normal (continuous linear Gaussian LCG) distribution. By analogy to (3), similar distribution holds for its conditional probability, which is expressed by (4)

$$P(t_{computed} \mid t_\mu\, ,\, t_\sigma^2\,) = N(t_\mu\, ,\, t_\sigma^2) \qquad (4)$$

and its causal model structure is shown in Figure 6. Here, the computed variable $t_{computed}$ is denoted by an ellipse with a double line boarder. The root nodes - real value denoted as $t\_real \equiv t_\mu$ and the sensor noise denoted as $t\_sigma \equiv t_\sigma^2$ - are modeled as discrete variables with uniform distributions for the purpose of independent treatment of the influence variables on the computed variable.

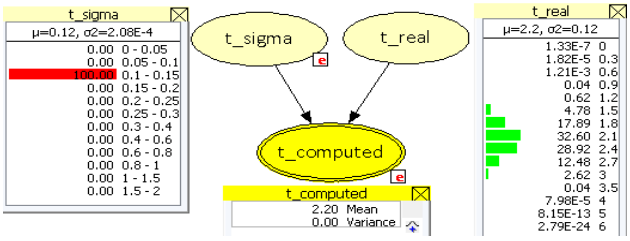


Figure 6. BN fragment, modeling uncertainties in a continuous variable $t_{computed}$, influenced by discrete variables t_μ and t_σ^2.

Similar principle of error estimation can be applied to any computed variable of interest for the BN modelling. The computed uncertainties complete the set of situation features, used for maneuver recognition.

In (Weidl *et al*, 2014), we proposed a number of modelling approaches to meet the automotive requirements on RAM and ROM memory size, and on computation time. These included besides the continuous nodes, also the use of function nodes as an alternative modeling of the sigmoid growth of probability for the hypotheses nodes; the use of expressions to specify the conditional probability distributions (CPDs) compactly and a divide-and-conquer approach to update of probability.

4.2 Modeling of driving behavior (hypotheses)

The ***lateral*** **relative dynamics (**Figure 4) is inferred from the actual lateral movement of a vehicle towards the lane marking. It is fused from the set of hypotheses $H_{1.1} \equiv$ {lateral evidence LE, actual movement trajectory TRAJ and free space FS, computed by the occupancy grid OCCGRID}. Here, the hypothesis LE fuses the vehicle's lateral offset to the lane marking (O_LAT) and its lateral speed (V_LAT) as shown in Figure 7. Its CPD is represented by a sigmoid function to expresses the growing probability for LE (and possible lane change) when the vehicle is coming closer to the lane marking (modeled by O_LAT_MEASURED) by growing lateral velocity (modeled by V_LAT_ MEASURED (see (Kaper *et al*, 2011, 2012, 2013) and for modeling optimization with continuous variables – see (Weidl *et al*, 2014). By analogy, the hypotheses TRAJ fuses: lateral acceleration (A_LAT), gear angle (vehicle's orientation in the lane) and the time-to-lane-crossing. For safety of lane change, $H_{1.1}$ checks available free space by assessing the risk of simultaneous occupancy of surrounding target cells (OCCGRID). This free space is inferred based on the times to enter and to leave the occupancy cells (Kaper *et al*, 2011, 2012, 2013).

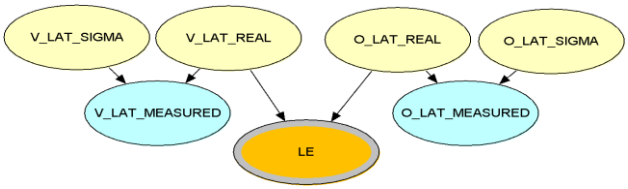


Figure 7. BN fragment modeling the hypothesis LE with discrete variables V_LAT_MEASURED and O_LAT_MEASURED.

The ***longitudinal*** **relative dynamics** (Figure 1, Figure 4) is fused from the set of hypotheses $H_{1.2} \equiv$ {longitudinal relative dynamics (RD) and its safety SAFE_RD}, see section 3. In the Original OOBN the BN-hypothesis *LE*, used for the evaluation of *LMC* (Figure 3, Figure 4), can recognize a maneuver only when the car approaches the lane marking. Hence, the intention of a driver to make a lane change cannot be detected with it. Therefore, we explore the longitudinal

relative dynamics (Figure 1) which is characterized by hypothesis "Relative Dynamics" (*REL_DYN* ≡ RD). Here we use the radar-measured features, characterizing the relation between a follower-vehicle and its front-vehicle on the same lane. The radar provides additional advantage of a longer view-horizon (up to 200 m) than the camera (up to 60 m). Since the hypotheses *REL_DYN* is contributing to the recognition of maneuver intention, it can be considered as "Maneuver Advice" and should be integrated in the higher abstraction OOBN layers, i.e. into the third layer in parallel with LMC and its output on "Maneuver Advice" further into the forth layer LC of the OOBN (Figure 3, Figure 4), since we use information on how fast the vehicles in front on the same lane are driving.

First, we apply the model for handling of uncertainties in measurements (described in section 4.1) - see layer 1, Figure 3. For simplicity, we will take two measured features (relative distance *X_REL_MEAS* and relative velocity *V_REL_MEAS* to the vehicle in front) and their variances σ^2 to improve the maneuver recognition time performance, i.e. earlier as with hypotheses $H_{1.1}$. The structure of the static BN-Model on relative dynamics is shown in Figure 8.

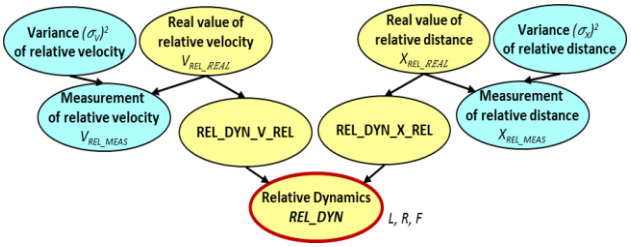


Figure 8. Static BN-Model for hypothesis "Relative Dynamics". Evidence nodes are coded with blue color; chance nodes – with yellow; decision hypothesis for maneuver L/R/F – with red border.

At the next abstraction level: $H_{1.1}$ contribute to the recognition of an event lane marking crossing by class LMC≡LANEMARKCROSS (where the classification node LMC is Boolean), which is reused at the next abstraction level for the recognition of LEFT and RIGHT lane marking crossing to infer on event LC≡lane change for each vehicle (OBJ1 and OBJ2). The OOBN model structure and parameters are based on domain knowledge and physical models, as described in details in (Kaper *et al*, 2011, 2012, 2013; Weidl *et al*, 2014, 2015). The initial OOBN parameterization has been specified qualitatively and quantitatively by use of expressions, like sigmoid functions or kinematics relations. The parameters have been initially hand tuned by domain experts to reflect expected lane change behavior for the conditional probability distributions of the hypotheses variables, which represent qualitatively a typical vehicle behavior at lateral and longitudinal relative dynamics with safety aspects.

4.3 Free Space Model for Safety

The CUTOUT and CUTIN maneuvers (Figure 1, Figure 2) can be considered as a lateral relative dynamics motion, since they represent a vehicle, performing a lateral movement towards the lane marking and relative to neighbor vehicles. In addition to the lateral relative dynamics, the longitudinal relative dynamics becomes essential for earlier recognition of maneuver intentions on lane change. It assumes, the analyzed vehicle aims to keep certain comfortable speed during its highway drive. It considers the longitudinal relative speed and relative distance to a vehicle driving in front on the same lane, Figure 1. The modeling principle of safety for the longitudinal relative dynamics is similar to the safety for the lateral dynamics, relative to the lane marking crossing (LMC), i.e., the hypothesis lateral Free Space (FS) in (Kasper, 2013). For safety, the longitudinal relative dynamics requires the check of available free space on the target lane to finish a maneuver or the suitability of a gap between two neighbor vehicles. This is performed by evaluating the safety features for longitudinal relative dynamics: "SAFE_RD/LEFT or RIGHT" (Figure 4). The driving praxis shows, that if a vehicle in front is slower, usually it is overtaken on the left, if free space is available (Figure 1). On the other hand, when a vehicle intends to leave the faster moving lane, it is slowing to change to the most-right or to the highway exit lane. Therefore, to ensure safety, it is necessary to estimate two features: the relative velocity and time to collision with vehicles on the target lane. These features are calculated as a relation to the nearest vehicles on the target lane, both behind and in front of the analyzed vehicle with a possible intention of a lane change. This safety mechanism is reflected in the fusion of the mentioned features, which are calculated for the left and right neighbor lanes. Figure 9 shows the BN fragment structure of the hypothesis "Safe_RD" for longitudinal relative dynamics. By analogy to the evaluation of the longitudinal relative dynamics to the front vehicle on the same lane (Tereshchenko, 2014; Weidl *et al*, 2015), a similar structure is used to model the relation to both the front and behind moving vehicle on the target lane (same fragment is used to evaluate both its left and right side).

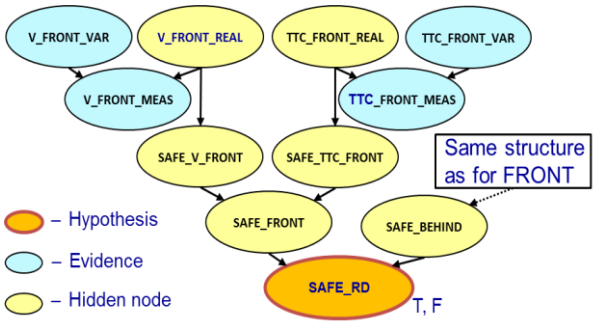


Figure 9. Safe hypothesis (SRD) for longitudinal Relative Dynamics.

The "Safe_RD" output nodes represent the interface nodes at the next layer of abstraction (Maneuver Advice ≡ MA, Figure 4).The evidence features are modeling the measurements (denoted as *_MEAS) with their uncertainty (variance denoted as *_VAR) which are assumed to have a Gaussian distribution. The distributions of nodes (V_FRONT_REAL, TTC_FRONT_REAL, etc.) are inferred based on the evidence. Nodes SAFE_V_FRONT and SAFE_TTC_FRONT are fused as an OR-relation in node SAFE_FRONT, i.e., a lane change to the target lane is safe only if at least one of the nodes has high probability. Thus, it models the relation between the "FRONT" input variables and the safety ahead on the neighbor target lane of the considered front vehicle. The CPD of SAFE_BEHIND is parameterized by analogy to SAFE_FRONT. Node SAFE_RD (evaluating the gap) combines the results from SAFE_FRONT and SAFE_BEHIND and is implemented as an "AND-relation", i.e. if both have a high probability for state "true", SAFE_RD will have also high probability for "true". However, if one of them is in state "false", SAFE_RD will have a high probability for "false".

4.4 Dynamic Models for Earlier Recognition

Here, we focus on the use of two-time slice dynamic Bayesian networks DBNs (2T-DBNs) to achieve earlier recognition of traffic maneuvers, see (Weidl *et al*, 2015). They are characterized by an initial model representing the initial joint distribution of the process and a transition probability distribution (TPD) representing a standard BN repeated over time. They satisfy both the first-order Markov assumption and the stationary assumption. Figure 10 shows the graphical structure of a 2T-DBN model for the hypothesis LE, while Figure 11 represents a DBN extension of hypothesis *REL_DYN*, with the hidden node $A_{REL\_REAL}(t)$, which was added for purposes as explained below. The TPDs between the time slices t and t+1 are assumed *conditional Gaussian* $N(\mu,\sigma^2)$. Here, since we do not have observations on the mean value μ, it is specified by physical models.

LE_DBN (Figure 10) is combining the real values of three lateral features: $O_{LAT\_REAL}(t)$, $V_{LAT\_REAL}(t)$ and $A_{LAT\_REAL}(t)$. When $O_{LAT\_REAL}(t)$ is steadily increasing and $V_{LAT\_REAL}(t)$ is high or increasing (requiring also $A_{LAT\_REAL}(t)$), their combination clearly indicates that the vehicle is leaving its lane. Note, that in (Kaper *et al*, 2011, 2012), A_{LAT_REAL} was included in hypothesis Trajectory (*TRAJ*) and not as part of *LE*. The TPDs for the *LE*-variables: $O_{LAT\_REAL}(t)$, $V_{LAT\_REAL}(t)$ and A_{LAT_REAL} are defined as shown in (5)-(7):

$$O_{LAT\_REAL}(t) \sim \mathcal{N}(O_{LAT\_REAL}(t\text{-}1) + V_{LAT\_REAL}(t\text{-}1) \cdot \Delta t,\ \sigma_{O\_LAT(t)}^2) \quad (5)$$

$$V_{LAT\_REAL}(t) \sim \mathcal{N}(V_{LAT\_REAL}(t\text{-}1) + A_{LAT\_REAL}(t\text{-}1) \cdot \Delta t,\ \sigma_{V\_LAT(t)}^2) \quad (6)$$

$$A_{LAT\_REAL}(t) \sim \mathcal{N}(A_{LAT\_REAL}(t\text{-}1),\ \sigma_{A\_LAT(t)}^2) \quad (7)$$

The time step Δt is the cycle time, i.e., 42 or 60 milliseconds depending on the camera used. The variances σ^2 are modeling the uncertainties of the variables. This dynamic extension incorporates the trend of real values, while their physics relations are represented as causal dependencies between time steps Δt. By analogy are defined the TPDs for the *REL_DYN* features: distance $X_{REL\_REAL}(t)$ and velocity $V_{REL\_REAL}(t)$ and the hidden variable relative acceleration $A_{REL\_REAL}(t)$.

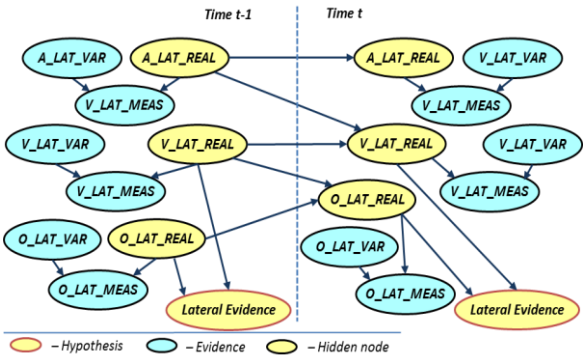


Figure 10. *LE_DBN*: 2T-DBN structure for the hypothesis *LE* (Lateral Evidence) for lateral Relative Dynamics towards the lane marking.

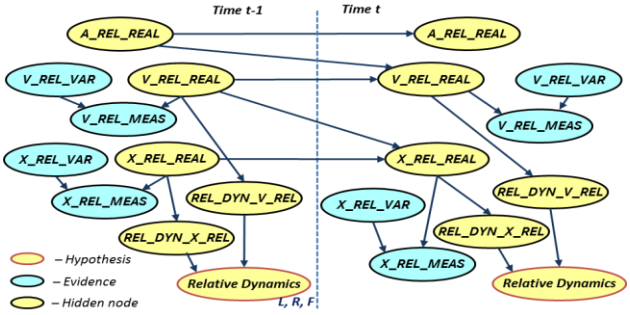


Figure 11. *REL_DYN_DBN*: The 2T-DBN structure for the hypothesis *REL_DYN* (Relative Dynamics) with *A_REL_REAL* as hidden node.

5 Approach

5.1 Combination of Methods

Our approach combines several methods to meet the deployment requirements on accuracy, less memory and faster inference. For resolving of programming paradigms, like efficient modelling and reuse of modeling fragments, we use OOBNs for information fusion from dynamic and/or static fragments. The accuracy requirement is reached by adding "special evidence" as described in section 5.3. Future study will also focus on parameters learning to further improve accuracy. In addition, to resolve the design paradigms for deployment on a prototype vehicle, i.e. to meet the requirements on computation time and memory, we have utilized parallelization of computations based on a divide-and-conquer strategy (D&C) (Weidl *et al*, 2014,

2015, 2017). This D&C parallelization splits the OOBN model into fragments and uses the posterior distribution of output nodes from the lower hierarchical fragments as likelihood over the corresponding input node at the next level of OOBN hierarchy, see (Weidl *et al*, 2017) for more details.

5.2 Modeling of the logical OOBN layers: LMC, MA, Lane Change Maneuvers and Driving Maneuvers (MNVR)

The network fragments, created to support the divide-and-conquer strategy to probability update in OOBN (Figure 3, Figure 4) are shown in Figure 12 and Figure 13. In Figure 12.A), LANEMARKCROSS (LMC) is the object class for *lateral* relative dynamics towards the lane marking. LMC represents the vehicle-to-lane-marking relation and is instantiated using the probabilities computed in the hypotheses TRAJ, LE, FS (OCCGRID for OBJ1 and OBJ2). In Figure 12.B), Maneuver Advice (MA) is the object class fusing the *longitudinal* relative dynamics REL_DYN (RD) between two vehicles driving on the same lane, and for safety evaluation on the left or right neighbor lanes, the available free space (Safe_RD) to a front and back vehicles, building the gap for completing a lane change (Figure 1). MA is instantiated using their probabilities.

The event class LANECHANGE (LC) is recognized by fusing LMC and MA (Figure 13). LC is instantiated by the probabilities, obtained from the hypothesis classes LMC towards left and right and from hypothesis class MA. The event class MNVR represents the vehicle-vehicle-vehicle LC-relation (denoted QMVT with 9 states, from all possible Left/Right/Follow LC-combinations of two objects) together with their relative lane-position to each other (denoted as POSDESCR with states: left, right, front). It infers the recognition of predicted maneuver, after instantiation by the probabilities, obtained from the two object classes LC.

5.3 Improving Accuracy by Special Evidence

Based on a performance analysis on sequences not included in the evaluation reported below, we have extended the OOBN models with measured/perceived variables representing *Special Evidence*. The evaluation has been performed with our statistical module and by additional visual examination. As typical for statistical classification, we use a confusion matrix to evaluate the classification results at each time step for all maneuver sequences. The corresponding maneuver state is classified (at each time step) as recognized (i.e. true positive TP, if corresponding to its reference data label) when its probability is bigger than 65%. This threshold value has been empirically derived in (Kasper, 2013). It has been derived from the confusion matrix and from a statistical evaluation of the probability of "false positives" for the Original OOBN on sequences not included in the evaluation reported below. Recall that all six maneuvers (Figure 2) are represented as a state of the MNVR variable, which has moreover one additional state for "DONTCARE" (i.e. pairs without any collision hazard).

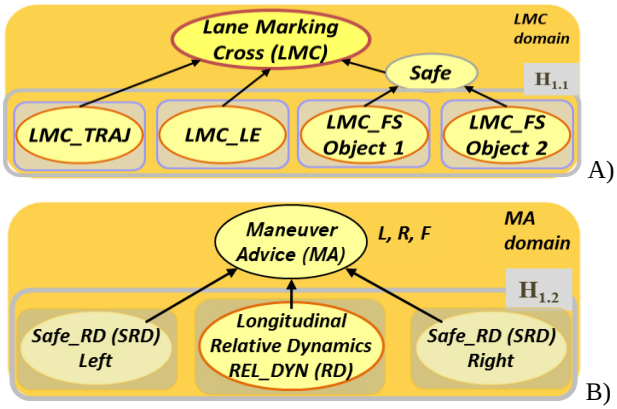


Figure 12. Classes: LANEMARKCROSS (LMC) and Maneuver Advice (MA).

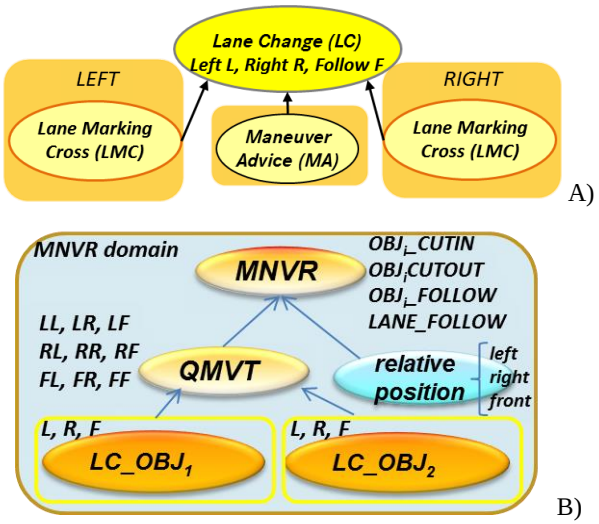


Figure 13. The object classes: LANECHANGE and Maneuver (denoted as MNVR)

We analyzed the results of the statisticsl evaluation and grouped the faults, based on causes with special attention to wrong classifications; and derived ideas for improvement of recognition performance for Lane-Follow and Object-Follow situations. We identified some *measured/percepted features* as "*special evidence* (SE_i)", which represent road topology and driving behavior in relation to other vehicles (which are present or not) on the same lane. These SE_i features extend the model of Figure 13.B) and are modeled with five blue nodes, while their influence on LC – with the yellow nodes, see Figure 14. Thus, they fuse the information for recognition of a lane change, where SE1 and SE2 consider the *lateral* and *longitudinal* relation between each pair of vehicles.

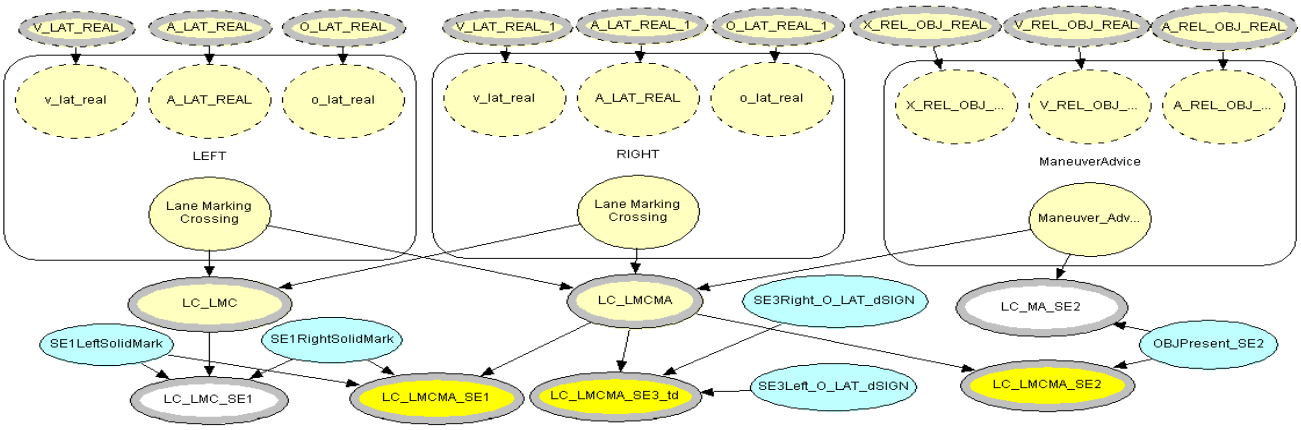


Figure 14. Information fusion at level LC of the extended OOBN model, including "special evidence" nodes to reduce false positive.

SE1: If one vehicle is approaching another vehicle moving on the most left (or most-right) lane, then no matter, if the longitudinal dynamics suggest LC to the left (or right) lane, this is not realizable due to natural lane boarders, i.e., there is no free space to execute any LC LEFT (or RIGHT) maneuver. This is incorporated in the road topology features *solid lane markings*: *SE1_LeftSolidMark, SE1_RightSolidMark.* It reduces the number of false positives (FP/wrongly classified) for OBJCUTOUT and the number of false negative (FN/not recognized) for FOLLOW maneuvers by 36% (from 11 to 7) even in the ORIG model.

SE2: If a vehicle is driving without any vehicle in front of it (possible even due to reduced perception reliability or out of sensor reach), then it has no reason to change the lane, unless an obstacle has been detected. This is incorporated in the model (Figure 14) by *SE2 OBJPresent (yes/no)*. This is represented by the logic rule: *If No Front Car detected, then both lateral and longitudinal dynamics classes* (Figure 4) - *LMC (LaneMarkCross) and REL_DYN - are set to Follow.*

SE3: Change of variable sign for lateral offset O_LAT. Due to the used coordinate system, the values of variables are positive only inside the current driving lane and change to negative, while changing to an adjacent lane. This is incorporated as features *SE3Right_O_LAT_dSIGN* and *SE3Left_O_LAT_dSIGN* in the extended structure of the BNs, Figure 14. The special evidence is present in the labeled data file and introduced in the models at the LC (Figure 14) and similarly at the maneuver MNVR (Figure 4) level.

5.4 Deployed classifiers

To study the effect of different model configurations on recognition, we have defined three static classifiers (ORIG; STATTR; STAT) and one dynamic classifier DBN; see Table 1. The ORIG classifier uses the Original OOBN (see Kaper *et al*, 2011, 2012; Kasper, 2013). "Y" shows which BN fragment is included in the corresponding classifier. All static classifiers use hypotheses LE and OCCGRID, while only ORIG and STATTR use TRAJ (Figure 4). The hypotheses "free space" FS ={OCCGRID and SRD} for the lateral and longitudinal relative dynamic respectively (Figure 4) remain static BN fragments for the purpose of satisfying the requirements on computation time and memory. The DBN fragments for the relative dynamics for the lateral and longitudinal motion are LE_DBN and REL_DYN_DBN. The developed static and DBN modelling fragments were generated as c-code and deployed by use of the divide-and-conquer (D&C) approach for probability update on the target Linux platform of the car – for details see (Weidl *et al*, 2017).

Table 1. Deployed classifiers on the Linux platform, see Figure 4.

	BN fragment				
Classifier	*LE (lateral relative dynamics)*	*TRAJ (trajectory)*	*OCCGRID OBJ1-OBJ2 (free space)*	*REL_DYN (longitudinal RD=relative dynamics)*	*SRD (free space/ safety)*
ORIG	Y	Y	Y	-	-
STATTR	Y	Y	Y	Y	Y
STAT	Y	-	Y	Y	Y
DBN	LE_DBN	-	Y	RD_DBN	Y

6 Evaluation and Analysis

Data sets used in the evaluation: The dataset has been acquired while driving in typical highway traffic. The raw data amounts to Terabytes. They are acquired by radar and stereo camera, which are fused to obtain the data objects with their characteristic features. In order to be able to analyze and use these data, they must be cleaned. The data preparation has involved: i) Visual examination of data quality for all collected sequences

in the prototype vehicle; ii) Statistical evaluation of all wrongly classified as well as not recognized maneuvers; iii) Data labelling and generation. Steps ii) and iii) have been automated. As a result, we have a total of 336 sequences consisting of 236 lane-change sequences and 100 lane-follow sequences. Quality measures of the recognition results have been selected for the performance evaluation of all developed classifiers. They include: the confusion matrix for the relevant (lane change) and irrelevant (follow or not present in the data) maneuvers; precision, recall; time gain for earlier recognition; and runtime performance of all classifiers. The statistical evaluation module establishes how big the time gain is relative to the labelled maneuver. The last is defined by the actual moment of lane marking crossing (LMC) by the midpoint of the car front bumper.

The confusion matrix for all deployed classifiers (Table 1) as performing on all evaluated driving sequences is shown in Table 2. The BN fragment TRAJ increases the accuracy, but the used gear angle is difficult to measure. The modeled special evidence has been successfully tested to contribute with a reduction of false positives (FP), thus increasing the accuracy. This has improved the performance on the Linux platform for deployment on the prototype car. We have made a proof, based on the performance of the ORIG classifier, by using all sequences and the statistical evaluation module, that the recognition accuracy of OOBN is not affected by the D&C approach and its implementation.

Table 3 shows the time gain for all maneuver classes for vehicle pairs. The average time gain for all deployed classifiers is about 1 second ahead of LMC. Moreover, dependent on the traffic situation and object perception, even earlier maneuver recognition is feasible. 36 of the tested sequences show earlier recognition with time gains of 1.5s – 6.72s (seconds). Test drives in real highway confirm, that traffic scenarios with "longitudinal relative dynamics" are recognized as a "need for lane change" before a vehicle is initiating a maneuver due to the recognition of a slower moving vehicle in front on the same lane. Therefore, the recognition by DBN classifier (visualized in Figure 15-17 with blue arrow) is earlier than the one by ORIG (visualized with red arrow). Figure 15 shows recognition of EGOCUTOUT, where DBN is 3.24s earlier than the actual lane marking crossing (LMC) and 2.46s earlier than ORIG. The recognition of OBJCUTOUT (Figure 16) is 4.62s earlier by DBN than LMC and 3.9s earlier than ORIG.

The driving sequence with the best recognition performance is shown in Figure 17, where OBJCUTIN maneuver is recognized by DBN 6.72s earlier than LMC and 5.88 s earlier than ORIG. The parallel D&C implementation has been deployed for all classifier alternatives (ORIG, STAT, STATTR, DBN, see Table 1.) on the automotive Linux platform for maneuver recognition. Table 4 shows between 22% and 40% gain in runtime performance for all deployed classifiers (using c-code on the Linux platform) due to the parallel D&C approach with "special evidence SE". A time performance of 1-4 milliseconds (ms) is still far from the target value of 0.15 milliseconds. One should note here, that these numbers are for the Linux prototype platform of the car and thus hardware dependent. The optimization of the parallel D&C method on a dedicated Linux computer allowed even better results with optimized time performance, coming very close to the initially set requirement of the target platform, see (Weidl *et al*, 2017).

Table 2. Performance comparison of the deployed classifiers with "special evidence" (index _SE). Column Nr.: 1: TP of OBJCUTIN; 2: TP of OBJ CUTOUT; 3: TP of EGOCUTIN; 4: TP of EGOCUTOUT; 5: TP (=as labeled); 6: FN (not recognized); 7: FP (wrong classified); 8: All (TP+FN) Maneuvers

Classifier	1	2	3	4	5	6	7	8	precision	recall
Label*	29	83	67	57	236	0	0	236		
ORIG	25	82	67	57	231	5	11	236	95.5%	97.9%
ORIG_SE	25	82	67	57	231	5	7	236	97.1%	97.9%
DBN_SE	23	80	67	56	226	10	19	236	92.2%	95.8%
STAT_SE	23	76	67	56	222	14	13	236	94.5%	94.1%
STATTR_SE	25	82	67	56	230	6	16	236	93.5%	97.5%

Table 3. Summary of evaluation results for time gain (negative time value means "maneuver prediction before crossing the lane marking") with all data for all deployed classifiers. 1: OBJCUTIN; 2: OBJCUTOUT; 3: EGOCUTIN; 4: EGOCUTOUT.

dt [s]	1	2	3	4	Avg. dt [s]
ORIG	-0.9950	-1.0129	-1.1461	-1.1156	-1.0749
ORIG_SE	-0.9979	-1.0129	-1.1936	-1.1945	-1.1089
DBN_SE	-0.8943	-0.9956	-1.1186	-1.3184	-1.1077
STAT_SE	-0.8977	-1.0161	-1.1192	-1.3092	-1.1089
STATTR_SE	-1.0150	-1.1246	-1.1545	-1.3500	-1.1763

Table 4. Runtime Performance.

Deployed classifier	Avg. Runtime [ms]	Deployed classifier by parallel D&C	Avg.Run time [ms]	Gain with parallel D&C [%]
ORIG	1.5989	D&C_ORIG	1.2508	21.8%
ORIG_SE	1.5989	D&C_ORIG_SE	1.0471	34.5%
DBN_SE	6.7405	D&C_DBN_SE	4.061	39.8%
STAT_SE	1.9734	D&C_STAT_SE	1.4606	26%
STATTR_SE	2.3143	D&C_STATTR_SE	1.5422	33.4%

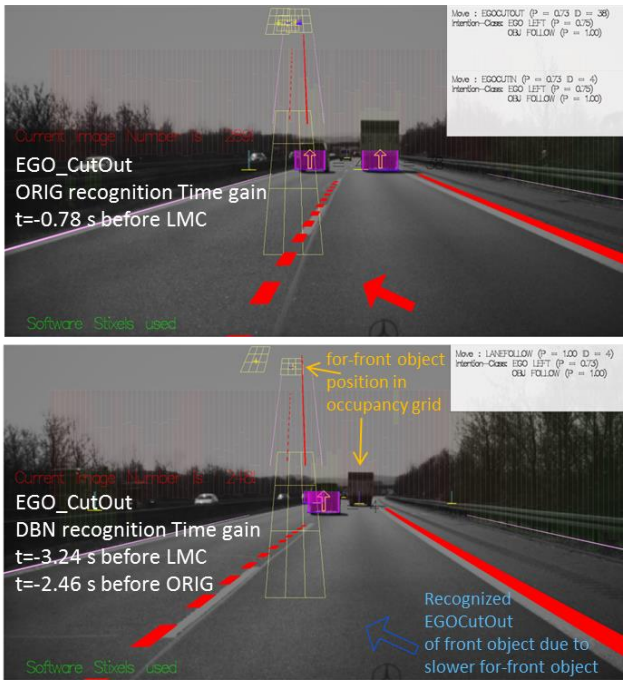


Figure 15. Highway demonstration with REL_DYN showing the classifier performance for EGOCUTOUT: DBN is 3.24s earlier than actual LaneMarkingCrossing (LMC) and 2.46s - than ORIG.

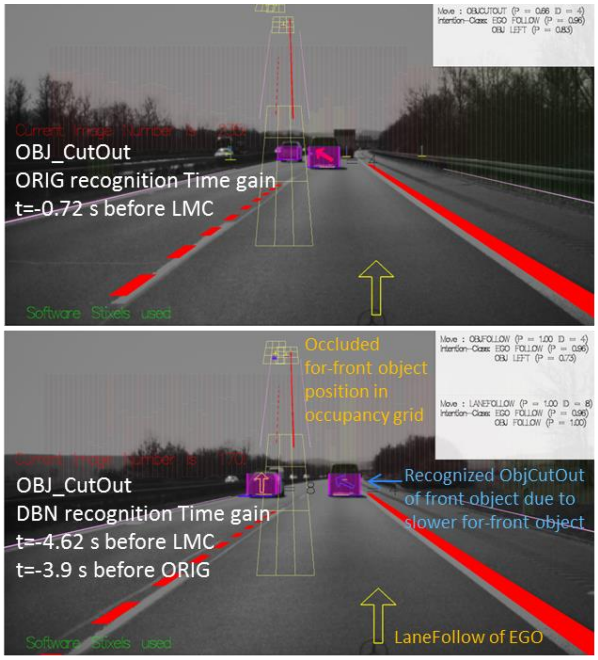


Figure 16. Highway demonstration with REL_DYN showing the classifier performance for OBJCUTOUT: DBN is 4.62s earlier than actual LMC and 3.9s earlier than ORIG.

7 Discussion of Results and Outlook

A Bayesian network has been designed and parameterized for lane change maneuvers. The advantage of our approach is that only measured features for lateral and longitudinal dynamics of the vehicles are necessary, without map data. The limitation is that lane markings are required to compute the features and some wrong classifications cannot be resolved for cases when the prediction horizon of a lane curvature does not reach the percepted front vehicle and thus the vehicle orientation inside the lane cannot be computed.

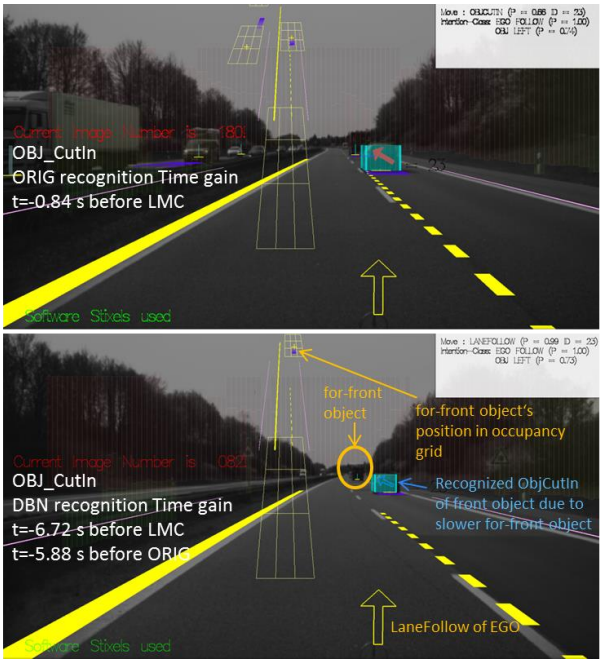


Figure 17. Highway demonstration with REL_D YN showing the classifier performance for OBJCUTIN: DBN is 6.72 s earlier than actual LMC and 5.88 s earlier than ORIG.

The introduced "special evidence" reflects road topology and vehicles relations, thus improving the recognition accuracy of lane-follow and reducing the false positives of lane-change maneuvers. The solution has been successfully tested for all classifiers. With the system deployment on the prototype vehicle, we have collected more data, which will be further divided to use for testing, and for learning of models' parameters of the lateral and longitudinal relative dynamics, together with their safety aspects. Here the hand tuned expressions will serve as an initial guess to improve further the recognition accuracy by use of machine learning techniques (Lauritzen, 1995), which have shown promising results.

In addition, we have analyzed the effect of parallelization of computations based on divide-and-conquer strategy (D&C). We describe in (Weidl *et al*, 2017) the implemented parallel D&C realization, allowing resolving the requirements on computation time (0.15 milliseconds) and memory for deployment on a prototype vehicle. This is an important step towards a scalable solution, meeting the hardware constraints of the automotive target platform. Future work will also focus on trend analysis for even more accurate and earlier maneuver recognition.

Acknowledgements

AMIDST (Analysis of Massive Data Streams) is a project, which has received funding from the European Union's 7th Framework Programme for research, technological development and demonstration under grant agreement no 619209.

References

N. Friedman and D. Koller. *Probabilistic Graphical Models: Principles and Techniques*, The MIT Press, 2009. ISBN-13: 978-0262013192.

D. Kasper, G. Weidl, T. Dang, G. Breuel, A. Tamke, and W. Rosenstiel. Object-oriented Bayesian networks for detection of lane change maneuvers. *IEEE Intelligent Vehicles Symposium (IV)*, 2011. doi: 10.1109/IVS.2011.5940468.

D. Kasper. *Erkennung von Fahrmanöovern mit object-orientierten Bayes-Netzen in Autobahnszenarien.* PhD-Thesis, Tübingen University Germany, 2013.

D. Kasper, G. Weidl, T. Dang, G. Breuel, A. Tamke, A. Wedel, and W. Rosenstiel. Object-oriented Bayesian networks for detection of lane change maneuvers. *IEEE Intelligent Transportation Systems Magazine*, vol.4, pp. 19–31, 2012. doi: 10.1109/MITS.2012.2203229.

U. B. Kjærulff and A. L. Madsen 2013, *Bayesian Networks and Influence Diagrams - A Guide to Construction and Analysis*. Springer. Second Edition, 2013. doi: 10.1007/978-1-4614-5104-4.

P. Kumar, M. Perrollaz, S. Lefevre and C. Laugier. Learning-Based Approach for Online Lane Change Intention Prediction. *IEEE Intelligent Vehicles Symposium (IV)*, Gold Coast, Australia, 2013.

S. L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, vol. 19, pp. 191-201, 1995. doi: 10.1016/0167-9473(93)E0056-A.

G. Li, S. E Li, Y. Liao, W. Wang, B. Cheng and F. Chen. Lane change maneuver recognition via vehicle state and driver operation signals — Results from naturalistic driving data. *IEEE Intelligent Vehicles Symposium (IV)*, pp.865-870, 2015. doi: 10.1109/IVS.2015.7225793.

A. Locken, H. Muller, W. Heuten and S. Boll. An experiment on ambient light patterns to support lane change decisions. *IEEE Intelligent Vehicles Symposium (IV)*, pp. 505-510, 2015. doi: 10.1109/IVS.2015.7225735.

M. Mori, K. Takenaka, T. Bando, T. Taniguchi, C. Miyajima and K. Takeda. Automatic lane change extraction based on temporal patterns of symbolized driving behavioral data. *IEEE Intelligent Vehicles Symposium (IV)*, pp. 976-981, 2015. doi: 10.1109/IVS.2015.7225811.

B. Morris, D. Anup and T. Mohan. Lane Change Intent Prediction for Driver Assistance: On-Road Design and Evaluation. *IEEE Intelligent Vehicles Symposium (IV)*, vol. IV, pp. 895-901, 2011. doi: 10.1109/IVS.2011.5940538.

R. K. Satzoda, P. Gunaratne and M. M. Trivedi. Drive quality analysis of lane change maneuvers for naturalistic driving studies. *IEEE Intelligent Vehicles Symposium (IV)*, pp. 654-659, 2015. doi: 10.1109/IVS.2015.7225759.

J. Schlechtriemen, A. Wedel, J. Hillenbrand, G. Breuel and K. D. Kuhnert. A lane change detection approach using feature ranking with maximized predictive power. *IEEE Intelligent Vehicles Symposium (IV)*, pp. 108-114, 2014. doi: 10.1109/IVS.2014.6856491.

V. Tereshchenko. *Relative object-object dynamics for earlier recognition of maneuvers in highway traffic.* Master's thesis, IAS, University of Stuttgart, Germany, 30.10.2014.

G. Weidl, A. L. Madsen, D. Kasper and G. Breuel. Optimizing Bayesian networks for recognition of driving maneuvers to meet the automotive requirements. *IEEE Multi-Conference on Systems and Control*, 2014. doi: 10.1109/ISIC.2014.6967630.

G. Weidl, A. L. Madsen, V. Tereshchenko, D. Kasper and G. Breuel. Early Recognition of Maneuvers in Highway Traffic. chapter: Symbolic and Quantitative Approaches to Reasoning with Uncertainty. *In Lecture Notes in Computer Science*, vol. 9161, pp. 529-540, 2015.

G. Weidl, A. L. Madsen, S. Wang, D. Kasper and M. Karlsen Early and accurate recognition of highway traffic maneuvers considering real world application: a novel framework using Bayesian networks. *IEEE Intelligent Transportation Systems Magazine*, accepted to appear in 2017.

F. Yan, L. Weber and A. Luedtke. Classifying driver's uncertainty about the distance gap at lane changing for developing trustworthy assistance systems. *IEEE Intelligent Vehicles Symposium (IV)*, pp. 1276-1281, 2015. doi: 10.1109/IVS.2015.7225858.

J. Zeisler, J. Cherepanov and V. Haltakov. A driving path based target object prediction. *IEEE Intelligent Vehicles Symposium (IV)*, pp. 316-321, 2015. doi: 10.1109/IVS.2015.7225705.

Y. Zhang and Q. Ji. Active and dynamic information fusion for multisensor systems with dynamic Bayesian networks. *IEEE Transactions on System&Cybernetics*, vol. 36/2, pp. 467–472, 2006. doi: 10.1109/TSMCB.2005.859081.

Monitoring Suspended Solids and Total Phosphorus in Finnish Rivers

Mauno Rönkkö and Okko Kauhanen[1] Jari Koskiaho, Niina Kotamäki and Teemu Näykki[2]
Markku Ohenoja and Esko Juuso[3] Maija Ojanen, Petri Koponen and Ville Kotovirta[4]

[1]Department of Environmental Science, University of Eastern Finland, Kuopio, Finland,
{mauno.ronkko,okko.kauhanen}@uef.fi,
[2]Finnish Environment Institute,Helsinki, Finland,
{jari.koskiaho,niina.kotamaki,teemu.naykki}@ymparisto.fi
[3]Control Engineering, Faculty of Technology, University of Oulu, Finland,
{markku.ohenoja,esko.juuso}@oulu.fi
[4]VTT Technical research Centre of Finland, VTT, Finland,
{maija.ojanen,petri.koponen,ville.kotovirta}@vtt.fi

Abstract

Monitoring of water quality should not be solely based on laboratory samples. Such activity, although producing reliable results, cannot provide an accurate enough temporal coverage for water quality monitoring. The Finnish Environment Institute, SYKE, has therefore established numerous online water monitoring stations that continuously monitor water quality. The problem with the automatic monitoring, however, is that the recorded values are not reliable as such and need to be subject to quality control and uncertainty estimation. Here, as the main contribution, we present a computational service that we have implemented to automate and integrate the water quality monitoring process. We also present a case study regarding the river Väänteenjoki and discuss the obtained uncertainty results and their implication.

Keywords: environmental measurements, quality control, uncertainty estimation, computational service, river

1 Introduction

The Finnish environmental authorities have regularly monitored water quality variables such as total suspended solids (TSS) and total phosphorus (TP) concentrations in major Finnish rivers since 1960s. The aim is to get general idea on the amounts of substances transported into the Finnish lakes and coastal areas and to detect possible trends. Presently, water quality and flow are monitored at 29 downstream monitoring stations of Finland's major rivers discharging into the Baltic Sea. There are also numerous inland river monitoring sites included, for instance, in the Finnish Eurowaternet (Niemi and Raateland, 2007). Annually, 13 to 22 water samples are taken and analyzed for TSS and TP (The Finnish Ministry of Environment, 2014) in addition to other substances.

SYKE has published quality recommendations for laboratories producing and delivering environmental monitoring data for registers of water quality in Finland (Näykki et al., 2013). Examples of the recommended limits of quantification (LOQ) and expanded measurement uncertainties (MU) are given in Table 1. The challenge in automating water quality monitoring is that the measurement data have not only significant seasonal variation but also erroneous values. Thus, without proper quality control and reliable uncertainty estimation, the data has little value. Thus far, this activity has been performed manually by environmental researchers.

Table 1. Recommended LOQ and MU (k=2) values for the measurement of TSS, TP and turbidity in monitoring of natural waters in Finland (Näykki et al., 2013).

Parameter	Unit	LOQ	Range	MU
TSS	mg/l	2	2-3	± 0.5
			> 3	± 20%
TP	µg/l	3	3-10	± 1.5
			> 10	± 15%
Turbidity	NTU (or FNU)	0.5	0.5-1	± 0.2
			>1	20 %

As a solution, we have implemented a computation service based on an Enterprise Service Bus Architecture. The service provides means for online quality control and integration of uncertainty estimation. It automates the sequence of operations performed earlier by the researchers; thus, providing more reliable results in a scalable and extensible manner. The service also enables reliable monitoring of the changes in water quality and flow over shorter periods of time.

2 Materials and Methods

The river Väänteenjoki was used here as the measurement site. A turbidity sensor was assembled at the Väänteenjoki site in the Karjaanjoki River Basin (2046 km^2) in southern Finland, shown in Figure 1, as a part of the sensor network

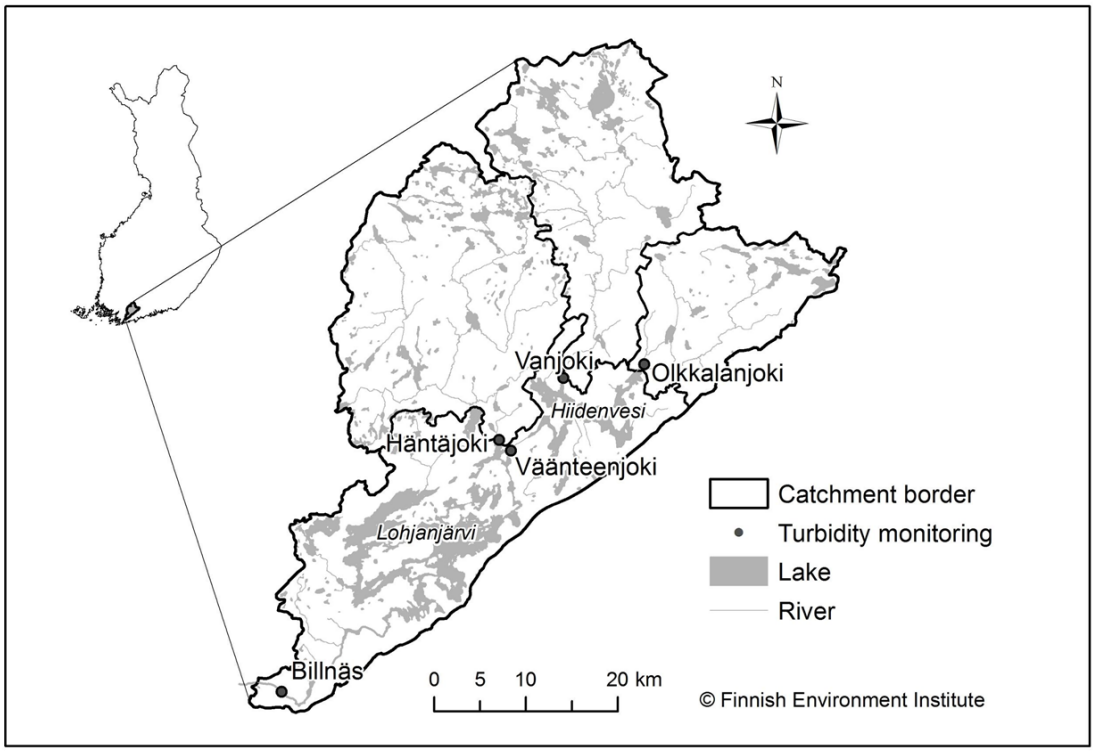


Figure 1. Location of the Väänteenjoki measurement site in the middle of the Karjaanjoki river basin.

established during the years 2007 and 2008 in the Soil-Weather project (Kotamäki et al., 2009). The river basin is mainly covered by forest (60%), the rest of the area being agricultural land (13%), lakes and rivers (12%), and population centers (9%). The Väänteenjoki measurement site is located between the two major lakes of the Karjaanjoki basin. From the lake Hiidenvesi (area 29 km2, mean depth 6.7 m), waters flow via the river Väänteenjoki into the lake Lohjanjärvi (area 92 km2, mean depth 12.7 m).

The monitoring process of the Väänteenjoki site is depicted in Figure 2. In the Väänteenjoki case, an OBS3+ sensor by Campbell Scientific was used. It emits a near-infrared light into the water, measures the light that scatters back from the suspended particles, and transforms this information into turbidity values (in Nephelometric Turbidity Units, NTU). The sensor collects and transmits the data to a server as SMS messages. The received messages are decoded as measurements with timestamps. The measurements are stored into a HydroTempo database maintained by SYKE.

Laboratory samples are collected once in a month. More frequent sampling can also be carried out during the high water runoff seasons. Laboratory turbidity measurement is carried out according to the international standard ISO 7027 (International Organization for Standardization, 1999). Turbidity is measured nephelometrically; the instrument measures the scattered light using the detector angle of 90 degrees from incident light. Instrument is calibrated with formazine standard solutions, and the turbidity of the tested water sample is expressed in Formazine Nephelometric Units (FNU). Note that FNU equal to NTU. TSS are determined at laboratory according to the standard EN 872 (European Committee for Standardization, 2005a), where water samples are filtered with GF/C glass fibre filter and dried at 105 ^{o}C. The mass of the residue retained on the filter is determined by weighing. Measurement of TP is based on standard EN ISO 15681-2 (European Committee for Standardization, 2005b) and recommendations of the analyzer manufacturer. Phosphorus is converted to orthophosphate by an acid-persulfate digestion prior measurement, where orthophosphate reacts in an acid solution containing molybdate and antimony forming an antimony phosphomolybdate complex. Reduction of the complex with ascorbic acid forms a strongly coloured molybdenium blue complex which is measured at the wavelength of 880 nm.

In a previous study (Koskiaho et al., 2015), turbidity recorded by an OBS3+ sensor at the Väänteenjoki site was calibrated against the turbidity determined from water samples taken near the sensor. Calibration equation, shown in Table 2, was determined according to linear regression between the values of the water samples and the simultaneous values recorded by the sensor. Because turbidity does not denote the content of substance in water, it cannot be directly used in calculations of material fluxes. Thus, correlations of turbidity with the concentrations of

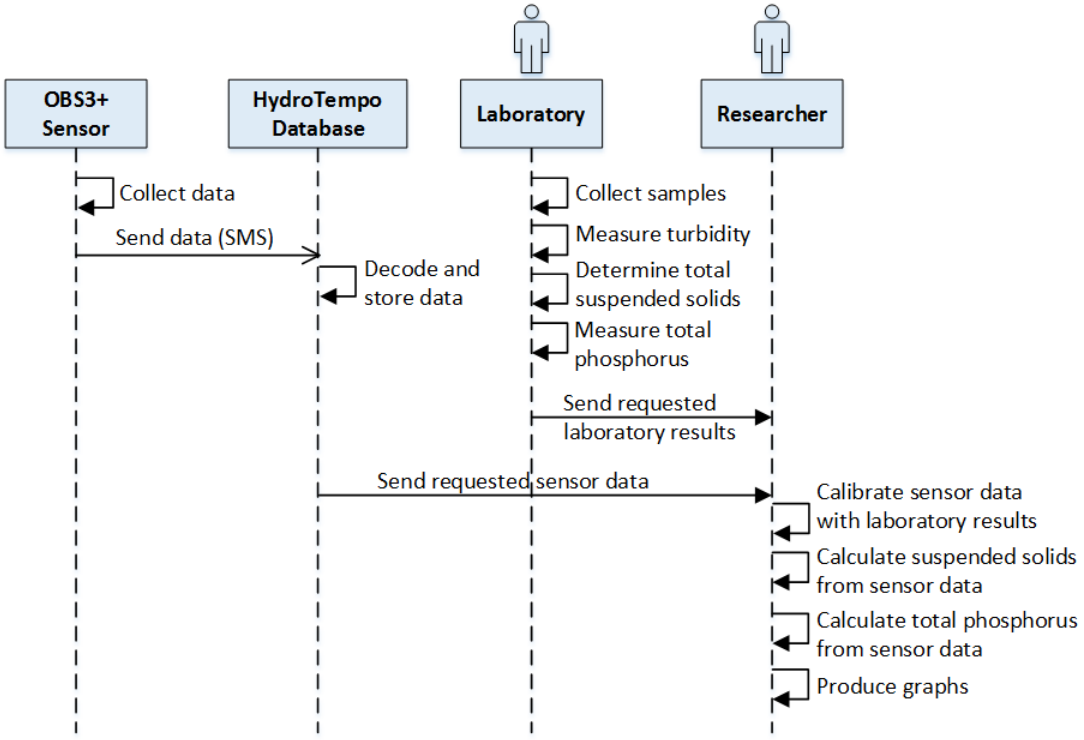


Figure 2. The monitoring process of the Väänteenjoki site.

TSS and TP were determined from the 2009–2012 water sample data collected at Väänteenjoki to convert the sensor-based, calibrated turbidity data to hourly concentrations of TSS and TP. The conversion equations based are presented in Table 2.

It should be noted that there may be events where turbidity jumps into the maximum (e.g. 250 NTU) and stays there for some hours. In some cases such events have lasted a day or two. The events could be caused by a storm followed by rapidly increasing water flow. In a previous study (Koskiaho et al., 2015), these events were manually checked, removed, and replaced with interpolated values before further processing. In (Koskiaho et al., 2015), however, measurement uncertainty was not estimated.

We have implemented a computational service that automates and integrates uncertainty estimation to the sequence of operations depicted earlier in Figure 2. The service architecture, shown in Figure 3, uses an Enterprise Service Bus (ESB) to connect different subsystems and to relay data between the subsystems. When the researcher inquires TSS for a specific river, ESB is used to acquire relevant measurement data and laboratory results. That data is passed to an online computational subservice. It computes TSS and TP for the river based on turbidity measurements. It also computes (combined) measurement uncertainty for the result.

In the computation, the calibration equation and related conversion equations are used as discussed earlier in Table 2. The computation results obtained by using these linear regression equations are then subjected to uncertainty estimation as discussed in (Ellison and Williams, 2012; Barwick, 2003). The results regarding the amount of TSS, including the linear regression equations, are then delivered to the researcher along with the (total) measurement uncertainty. The measurement uncertainty is a reliability estimate and can be used to compare the measurement results among each other or with reference values (JCGM, 2008).

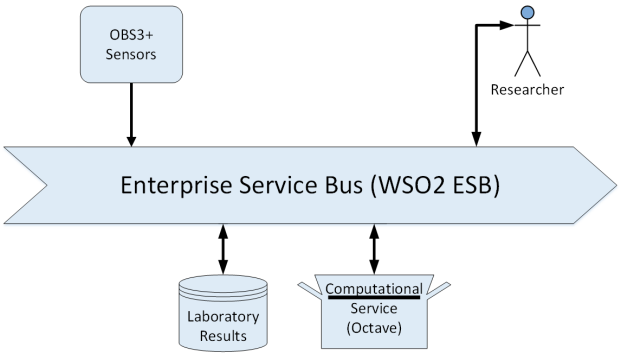


Figure 3. ESB based architecture implementing the service for computing TSS and TP along with measurement uncertainty estimation.

The combined measurement uncertainty includes contribution from the laboratory reference measurement and the performance of the measurement models, i.e. how well the measured values fit in the model, and how repeatable the online measurements are. The laboratory reference uncertainty is estimated according to the Nordtest approach (Magnusson et al., 2012) and ISO standard 11352 (International Organization for Standardization, 1012), where

Table 2. Calibration equations and coefficients of determination (r^2) derived from the relation between the sensor recordings and water samples, and equations to convert the calibrated turbidity to TSS and TP concentrations.

Calibration equation	Conversion equation for the concentration of	
	TSS [mg/l]	TP [µg/l]
y = 2.8 + 2.12x; r^2=0.86	z = 1.6 + 0.78y; r^2=0.74	z = 14.3 + 1.76y; r^2=0.82

where x = recorded turbidity (NTU, raw data)
y = calibrated turbidity (NTU, final data for further calculations)
z = TSS or TP concentration determined from clibrated turbidity

Table 3. Uncertainty budget for the turbidity, TP and TSS concentration measurements.

	Turbidity (25 FNU)	TP (60 µg/l)	TSS (20 mg/l)
Laboratory analysis	10 %	7.5%	7.5%
Turbidity sensor measurement model	7.8%	7.8%	7.8%
Phosphorus/solids measurement model		7.7%	11.9%
Combined standard uncertainty	12.8%	16.4%	17.4%
Expanded uncertainty (k=2)	25.6%	32.9%	38.0%

combined standard uncertainty consists of two main components: the within-laboratory reproducibility and the uncertainty due to possible bias. The quality recommendation for reporting measurement results and uncertainties does not include the uncertainty due to sampling (Näykki et al., 2013). The uncertainty of the linear regression measurement models, u_{c_0} , is estimated with help of (Ellison and Williams, 2012):

$$u_{c_0} = \frac{S}{B_1}\sqrt{\frac{1}{p}+\frac{1}{n}+\frac{(c_0-\bar{c})^2}{S_{xx}}} \qquad (1)$$

where S is the residual standard deviation of the measurement model, B_1 is the slope of the measurement model, p is the number of replicate measurements in the determination of the concentration c_0, n is the number of replicate measurements in the determination of the mean calibration value $\bar{c}$, and S_{xx} is the standard deviation of the measured values and the mean calibration value. The uncertainty related to the repeatability of the measurements is included in Equation (1).

3 Results and Discussion

Hourly time series of the calibrated turbidity measured by sensors (curves) together with turbidity analyzed from water samples (dots) are presented in Figure 4. As suggested by the equality of sampled and sensor-based values, calibration of the sensors was successful.

The turbidity curve of the river Väänteenjoki, also shown in Figure 4, differed strongly from those of the other measurement sites dealt with in (Koskiaho et al., 2015), which showed clearly sharper general form and higher peak values. The difference was claimed to be a consequence of the retention effect of the lake Hiidenvesi and the close proximity of the Väänteenjoki measurement site to the lake. In other measurement sites, the distance to the upstream lakes was much longer, or the lakes were small.

The uncertainty components are listed in Table 3. In this example, a typical value for the turbidity is 25 FNU, for the TP is 60 μg/l, and for TSS is 20 mg/l. The standard uncertainty (k=1) of the laboratory analysis is 10% for the turbidity in the range $>$1 FNU, and 7.5% for the TP and TSS in the ranges of $>$10 μg/l and $>$3 mg/l, respectively. Using Equation (1), we obtain the standard uncertainty estimates for the turbidity sensor calibration, 7.8%, and for the TP and TSS measurement models, 7.7% and 12%, respectively. The TP and TSS contents are determined based on turbidity, so the calibration of the turbidity sensor must be included. The combined standard uncertainty is calculated as a square sum of the individual, uncorrelated components (JCGM, 2008). Figure 5 demonstrates the expanded uncertainty (in content) as a function of measured turbidity, TP and TSS. The measured values presented are between the minimum and maximum values of the data set of our example.

It can be seen that all the uncertainty components, laboratory analysis and the measurement models, are of the order of the same magnitude. The linear regression models are based on the whole dataset 2009 – 2012. In this case, there is no significant difference in terms of uncertainty whether the model is fitted based on the whole dataset or on yearly basis.

Using the presented measurement models practically doubles the measurement uncertainty as compared with the laboratory analysis, but on the other hand the online monitoring provides more representative and frequent information in the water quality. In Figure 6, this is illustrated by presenting the relative standard error of the mean for TP in different variation levels, calculated based on the two different monitoring processes and their measurement uncertainties presented in Table 3.

The more frequent data of online monitoring (24 samples per day) enables following the trends in daily basis with low estimation error, whereas infrequent laboratory

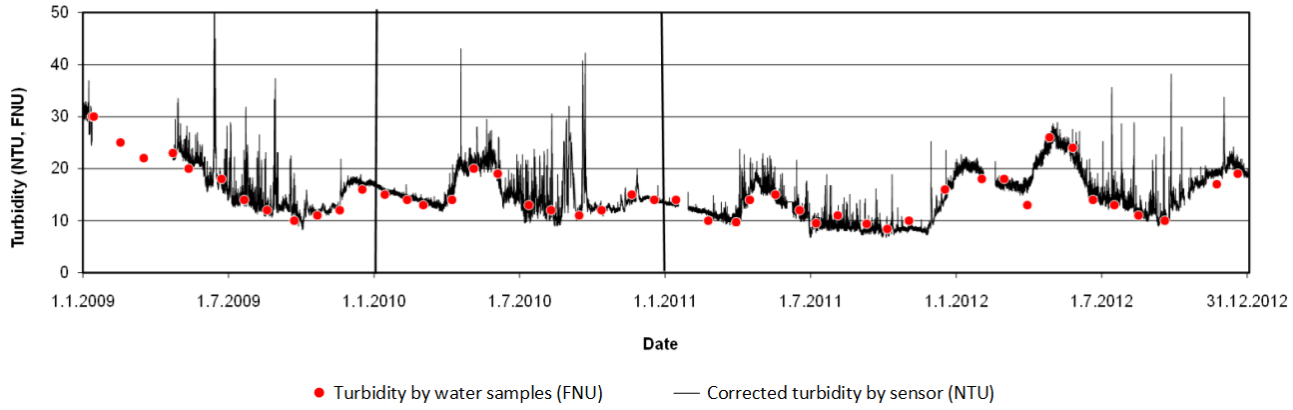


Figure 4. Hourly time series of the corrected turbidity measured by sensors (NTU) together with turbidity analyzed from water samples (FNU) in Väänteenjoki measurement site.

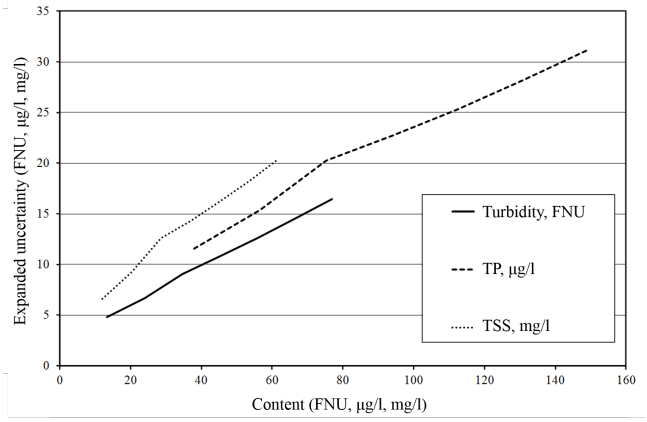


Figure 5. Expanded uncertainties of turbidity (FNU, solid line), TP(μg/l, dashed line) and TSS (mg/l, dotted line) as a function of content.

measurements (1 sample per month) is only suitable for estimating yearly averages with a comparable estimation error.

As compared with the quality recommendations in Table 1, the uncertainty of the laboratory analysis is equal to the recommendation for turbidity and TP. For TSS, the quality recommendation is 20%, and in this case 15%.

For the measurement models, the total number of measurement data points is 45 for turbidity and 46 for TP and TSS. The coefficients of determination for the measurement model equations are 0.86, 0.82 and 0.74 for the turbidity, TP and TSS measurement models, respectively. The coefficient of determination of 0.90 would result in uncertainties of 6.5%, 6.0% and 7.7% for turbidity, TP, and TSS, respectively. With the present laboratory analysis uncertainties, these uncertainties would produce respective expanded uncertainties of 24%, 31% and 32%. If the coefficient of determination is 0.95, the turbidity sensor measurement model would result in 4.5% standard uncertainty. This would result in 22% expanded uncertainty in turbidity, 29% in TP, and 31% in TSS measurement model.

However, the coefficient of determination of 0.90 would be obtainable by using an outlier detection algorithm. This would require removal of two most deviating points from turbidity model, five from TP model and ten points from the TSS model. Even the coefficient of determination of 0.95 is obtainable, at least with turbidity measurement model. This would require removing 11 most deviating measurement points from the model. Reducing the size of the data set down to 34 points does not increase the uncertainty of the model significantly.

4 Conclusions

A computational service for monitoring TSS and TP was introduced in this study. The service also integrates measurement uncertainty estimation and delivers monitoring results with uncertainty information. The service was implemented by using a scalable and extensible architecture based on the use of an Enterprise Service Bus.

The monitoring results obtained for the river Väänteenjoki were presented and discussed. The results indicate that online monitoring does not fully fit within the recommended limits of quantification. The advantage of online monitoring, however, is that it is a continuous activity and it supports monitoring of quick events and changes in trends. This is something that would not be feasible by using laboratory sampling only. Thus, in together with laboratory sampling online monitoring provides a more accurate situation picture of the state and quality of Finnish rivers.

Acknowledgements.

We thank the Finnish Funding Agency for Technology and Innovation (TEKES) for funding this research. This work has been carried out in TEKES funded CLEEN SHOK programme "Measurement, Monitoring and Environmental Assessment".

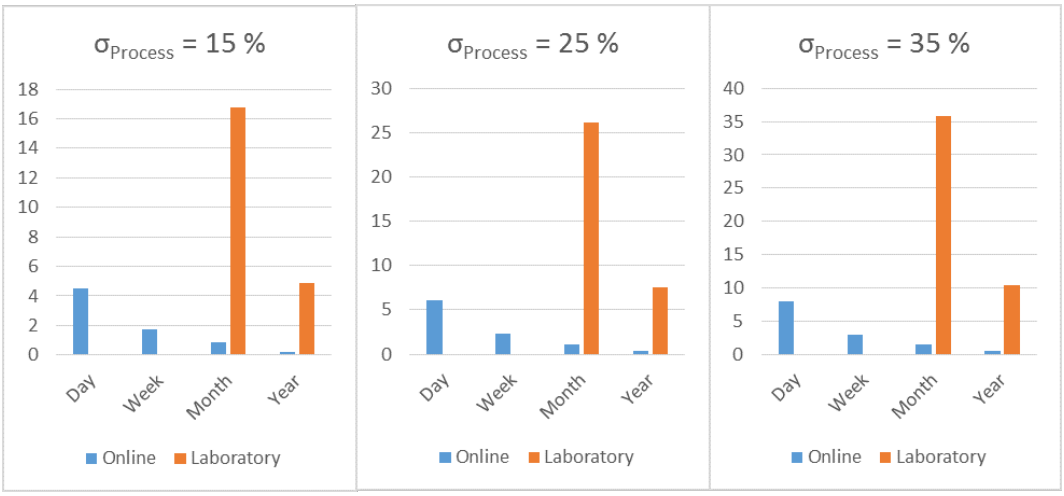


Figure 6. Relative standard error of the mean (SE) with the different monitoring processes and different observation periods calculated as $(\%) = 100\sqrt{(\sigma_p^2 + \sigma_M^2)/n}$.

References

V. Barwick. Preparation of calibration curves: A guide to best practice, 2003.

S. L. R. Ellison and A. Williams. Eurachem/citac guide quantifying uncertainty in analytical measurement, third edition, 2012.

European Committee for Standardization. EN 872. water quality - determination of suspended solids, 2005a.

European Committee for Standardization. EN ISO 15681-2. water quality - determination of orthophosphate and total phosphorus contents by flow analysis (FIA and CFA). part 2: Method by continuous flow analysis (CFA), 2005b.

International Organization for Standardization. ISO 11352:2012, water quality. estimation of measurement uncertainty based on validation and quality control data, 1012.

International Organization for Standardization. ISO 7027. water quality - determination of turbidity, 1999.

JCGM. Evaluation of measurement data - guide to the expression of uncertainty in measurement (jcgm 100:2008, gum 1995 with minor corrections), 2008.

J. Koskiaho, S. Tattari, and E. Röman. Suspended solids and total phosphorus loads and their spatial differences in a lake-rich river basin as determined by automatic monitoring network. *Environmental monitoring and assessment*, 187(4):1–12, 2015.

N. Kotamäki, S. Thessler, J. Koskiaho, A. O. Hannukkala, H. Huitu, T. Huttula, J. Havento, and M. Järvenpää. Wireless in-situ sensor network for agriculture and water monitoring on a river basin scale in Southern Finland: Evaluation from a data user's perspective. *Sensors*, 9(4):2862–2883, 2009.

B. Magnusson, T. Näykki, H. Hovind, and M. Krysell. Nordtest technical report 537: Handbook for the calculation of measurement uncertainty in environmental laboratories, 3.1 edn., 2012.

T. Näykki, H. Kyröläinen, A. Witick, I. Mäkinen, R. Pehkonen, T. Väisänen, P. Sainio, and M. Luotola. Quality recommendations for data entered into the environmental administration's water quality registers: Quantification limits, measurement uncertainties, storage times and methods associated with analytes determined from waters (in Fnnish), environmental administration guidelines 4/2013, 2013. URL `https://helda.helsinki.fi/handle/10138/40920`.

J. Niemi and A. Raateland. River water quality in the Finnish eurowaternet. *Boreal Environment Research*, 12:571–584, 2007.

The Finnish Ministry of Environment. Suomen merenhoidon seurantakäsikirja (marine management monitoring handbook of Finland), 2014. URL `www.ymparisto.fi/download/noname/$%$7BD36B07E3-30F7-4F48-8B55-D809AC74FA8B$%$7D/98219`.

Artificial Neural Networks Application in Intraocular Lens Power Calculation

Martin Sramka[1] Alzbeta Vlachynska[2]

[1]Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic, sramkma2@fel.cvut.cz
[2]Faculty of Applied Informatics, Tomas Bata University in Zlin, Zlin, Czech Repubic, vlachynska@fai.utb.cz

Abstract

This article deals with intra-ocular lens (IOL) power calculations during the cataract surgery. At present, IOL power calculated by formulas is usually able to provide acceptable results for the majority of the patients. The problem appears when any of input parameters have the value which is not normal in population distribution. Then the patient post-operative refraction result can inconsiderable deviate from intended target. This work describes approach how to preoperatively indicate which samples of a patient could be problematic in accurate IOL calculations by classification of Artificial Neural Networks (ANN). Small and long eyes are used to test the ability of ANN to classify input samples which are taken from pre-operative measurements to several groups which represent probable post-operative result. In our experiment, ANN classifies samples into two groups. The first group is for data samples with a probable result in positive ranges of diopter and second group is for negative ranges. The accuracy of ANN, in this case, is 94.1 %.

Keywords: intra-ocular lens (IOL) power calculation, artificial neural networks (ANN), cataract surgery, refraction result

1 Introduction

A cataract is present when the transparency of the eye lens is reduced to the point that the patient's vision is impaired. According to the latest assessment of World Health Organization, cataract is responsible for 51 % of world blindness which represents about 20 million people (2010). Fortunately, it can be surgically removed. Natural lens of the eye that has developed an opacification is replaced with an artificial IOL. Choosing the appropriate IOL power is a major determinant of patient satisfaction with cataract surgery. There are 3 main factors: accurate measurements (biometry), selecting appropriate calculations (formulas), and assessing the patient's needs and expectations to determine the postoperative refractive target (clinical considerations) (Henderson et al., 2014).

When the human lens is replaced with an intra-ocular lens, the optical status becomes a two-lens system (cornea and intra-ocular lens) projecting an image onto the fovea. The distance (X) between the two lenses affects the refraction as does the distance (Y) between the two-lens system and the fovea. X is defined as the distance from the anterior surface (vertex) of the cornea, which curvature is described by keratometry (K), to the effective principle plane of the intra-ocular lens in the visual axis. Y is defined as the distance from the principal plane of the intra-ocular lens to the photoreceptors of the fovea in the visual axis. It is easy to see in Figure 1 that X + Y is equal to the visual axis, the axial length of the eye (AL). Therefore, knowing X and A will allow the calculation of Y (Y = AL - X). Also to calculate the intra-ocular lens power (P), we must know the vergence of the light rays entering the cornea (refractive error (R)). For emmetropia, R is zero. The relationship of these factors (X, Y (AL-X), P, K, R) is such that a formula can be written to describe it. Knowing the values of any four of these variables will allow for the calculation of the fifth (Roger and David, 2010).

1.1 Calculation formulas

For appropriate intra-ocular lens power selection, the mathematical computing formulas are used. These formulas have been developed approximately from the sixties of the last century (Kuchynka, 2007). Over time, some trends have emerged regarding which formulas to use in general categories of patients:

- <22 mm: Hoffer Q
- 22-23 mm: Hoffer Q or Holladay 1
- 24-26 mm: Holladay 1
- >26 mm: SRK/T or Holladay 2

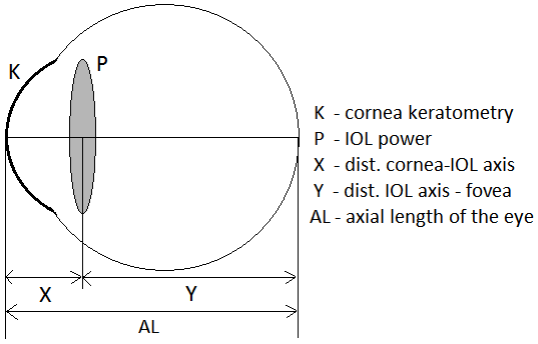


Figure 1. Ocular dimesions.

It is essential to use the appropriate power calculation constant (A constant, ACD-constant, surgeon factor) specified by the intra-ocular lens manufacturer for the specific formula, chosen intra-ocular lens style, and personalized as warranted by the surgeon (Henderson et al., 2014). These formulas usually work very well for the majority of the patients, but the problem may appear if any of input parameter has a value which is not normal in population distribution.

1.2 Calculation of Intra-Ocular Lens for Non-Normal Eyes

Many results and ways how to solve the problem of intra-ocular lens power calculation can be found in the present literature. The problem of accurate calculation is quite complex and depending on many pre and post-operative factors. Therefore needs to be divided into many parts.

One of many ways how to calculate intra-ocular lens power can be seen in (Abulafia et al., 2015). The accuracy of Holladay 1, SRK/T, Hoffer Q, Haigis, Barrett Universal II, Holladay 2, and Olsen formulas for eyes axial length longer than 26.0 mm is provided. SRK/T, Hoffer Q, Haigis, Barrett Universal II, Holladay 2, and Olsen methods are having a prediction error of $\pm$0.5 D in at least 71 % of eyes and $\pm$1.0 D in 93 % of eyes.

A calculation for 53 eyes of 36 patients with axial length more than 27.0 mm by the IOL Master is evaluated in (Bang et al., 2011) for the Holladay 1, Holladay 2, SRK/T, Hoffer Q and Haigis formulas. For eyes longer than 27.0 mm the Haigis formula is found to be most accurate followed by SRK/T, Holladay 2, Holladay 1 and in the last place Hoffer Q. All formulas predicted more myopic outcome than was the real result of the surgery.

Refractive outcomes for small eyes and calculation with Hoffer Q, Holladay 1, Holladay 2, Haigis, SRK-T, and SRK-II are observed in (Carifi et al., 2015). The Hoffer Q formula provides best refractive outcomes of which 39 %, 61 %, and 89 % of the eyes had a final refraction within $\pm$0.5 D, $\pm$1.0 D, and $\pm$2.0 D of the target, respectively.

The accuracy of Hoffer Q and Haigis formula according to anterior chamber depth in small eyes is evaluated in (Eom et al., 2014). 75 eyes of 75 patients with axial length less than 22.0 mm is included in the study. The difference between the predicted refractive errors of the Hoffer Q and Haigis formula increased as ACD decreased in short eyes. No significant difference is found when anterior chamber depth is longer than 2.4 mm.

Predictability of intra-ocular lens power calculation using Carl-Zeiss IOL Master and applanation ultrasound using SRK/T, SRK II, Holladay 1 and Haigis with an axial length longer than 25.0 mm is evaluated in (Wang et al., 2008). The mean axial length was significantly longer than in case of applanation ultrasound. The mean average errors calculated by the SRK/T, SRK II, and Holladay 1 formulas were comparable between both methods of measurement. The best results were provided by the IOL Master data in combination with the Haigis formula.

Refractive prediction of Holladay 1, Hoffer Q, Haigis and SRK/T intraocular lens power calculation formulas for eyes longer than 25.0 mm is evaluated and method for axial length optimization is proposed in (Wang et al., 2011). Refractive prediction errors with the Holladay 1, Haigis, SRK/T, and Hoffer Q formulas were evaluated in consecutive cases. Eyes were randomized to a group used to develop the method of optimizing AL by back-calculation or a group used for validation. Further validation was performed in two additional datasets. The proposed method of optimizing AL significantly reduced the percentage of long eyes with a hyperopic outcome. Updated optimizing AL formulas by combining all eyes from the two study centers are proposed.

Refractive outcomes of Haigis–L formula for calculation intraocular lens power in Asian eyes with axial lengths longer than 25.0 mm that had a previous myopic laser in situ keratomileusis or photorefractive keratectomy are evaluated in (Wong et al., 2015). The predictability of being within $\pm$0.5 D and $\pm$1.0 D of the target was 35.7 % and 63.1 %, respectively. 31.6 % and 60.5 %, respectively, in eyes with an AL less than 27.0 mm; and 39.1 % and 65.2 %, respectively, in eyes with an AL of 27.0 mm or longer.

Next interesting way how to calculate an intra-ocular lens power is provided by (Clarke and Burmeister, 1997). The accuracy of trained artificial neural network and Holladay 1 formula is compared. In 72.5 % of cases for artificial neural network and in 50.0 % of cases for Holladay 1 formula an error of less than $\pm$0.75 D is achieved.

2 Back ground of studies

As was described; many ways how to calculate intra-ocular lens power is being used at present. Many algorithms and calculation formulas have its own accuracy limits for example in the calculation for the unusual cases of eyes - eyes with long or short axial length or keratometry.

In the field of cataract surgery as well as in all health-care related fields patients demands and expectations to provided care are continuously increasing. Together with increasing prevalence of cataract surgery, this could be one of the main motivating factors to provide best possible postoperative refraction results to if possible the greatest amount of patients. Another factor which is no less important could be the economic side of re-operation or following refractive correction of the patient which has implanted bad calculated intra-ocular lens.

As was written in the previous section of this article. Our algorithm for intra-ocular lens power estimation is based on Artificial Neural Networks which are used as a classifier.

3 Artificial Neural Network

Artificial Neural Networks dominate by ability in immediate pattern recognition of input/output relations. This

differs ANN from expert systems which achieve excellent results in the sequence of logical operations and fuzzy logic methods and are characterized by the ability represent knowledge (Tuckova, 2009).

Artificial Neural Networks are inspired by biological neural networks. This property in some way determines that the Artificial Neural Network should be capable to behave well or at least like their biological patterns. Knowledge of Artificial Neural Network is stored in relations between neurons. These relations are strengthened during the learning process or penalized when the learning does not lead to better results. More about Artifical Neural Networks can be found in (Tuckova, 2009).

Our algorithm for intraocular lens power estimation is based on Artificial Neural Network which is used as a classifier. The base principal of our access is following:

1. Multi-layer Artificial Neural Network classifies input matrix compound from pre-operatively measured K, ACD and AL into several groups.
2. Each group has its own estimated post-operative refraction outcome.
3. Intra-ocular lens power is calculated using standard SRK/T formula.
4. Intra-ocular lens power calculated by SRK/T formula can be corrected by surgeonâĂŹs decision based on refractive outcome estimated by classification of Artificial Neural Network.

3.1 Real data

For the research purposes, specific patientâĂŹs data had to be collected from clinic database, then cleaned up and preprocessed.

3.1.1 Collected Data

- Implanted IOL manufacturer and type
- Pre operative ïňĆat meridian of the cornea - K1 [D]
- Pre operative steep meridian of the cornea - K2 [D]
- Pre operative anterior chamber depth - ACD [mm]
- Pre operative axial length of the eye - AL [mm]
- Implanted IOL power - P [D]
- Implanted IOL A constant - A
- Subjective post operative sphere - SfS [D]
- Subjective post operative cylinder - CyS [D]
- Subjective post operative spherical equivalent - SfES [D]
- Objective post operative sphere - SfO [D]
- Objective post operative cylinder - CyO [D]
- Objective post operative spherical equivalent - SfEO [D]
- Eye

3.1.2 Specification of Patients Data Selected for Collection

- Patient undergoing cataract surgery
- Indicated for monofocal intra-ocular lens
- Data from beginning of January 2012 till end of July 2015
- Both eyes
- Calculated for emetropia
- Pre operative ïňĆat meridian of the cornea between 30 and 55 diopters
- Pre operative steep meridian of the cornea between 30 and 55 diopters
- Pre operative anterior chamber depth between 1 and 5 millimeters
- Axial length of the eye between 15 and 21 or between 25 to 35 millimeters
- Post operative sphere between -10 and 10 diopters
- Post operative cylinder between -10 and 10 diopters
- No cases with previous corneal refractive surgery

3.1.3 Specification of Preprocessing Parameters

- SfCalc: post operative sphere calculated by SRK/T from K1, K2, AL, P and A constant.
- RDS: difference between SfCalc and SfS.

$$RDS = SfCalc - SfS \tag{1}$$

- RDO: difference between SfCalc and SfO.

$$RDO = SfCalc - SfO \tag{2}$$

- K: mean keratometry calculated from K1 and K2.

$$K = \frac{K1 + K2}{2} \tag{3}$$

- InputVector: matrix which is used as input for Artificial Neural Networks and is composed from AL, ACD and K vectors, which are normalized between 0 and 1.

$$ALn = \frac{AL - min(AL)}{max(AL) - min(AL)} \tag{4}$$

$$ACDn = \frac{ACD - min(ACD)}{max(ACD) - min(ACD)} \quad (5)$$

$$Kn = \frac{K - min(K)}{max(K) - min(K)} \quad (6)$$

- TargetVector: this variable separates InputVector into several groups and is used for Artificial Neural Network training and testing. Ranges of these groups are described in results section.

On the Figures 2, 3, 4 there can be seen dependence of RDS between AL, Kmean and ACD of tested data. Most significant dependency can be seen on Figure 2 where RDS grows with decreasing AL and conversely. Some other trends can be also find in Figure 3 or Figure 4.

These multifactorial dependencies are the main reason why we chose the Artificial Neural Networks as our main decision algorithm for intra-ocular lens power estimation improvement.

4 Results

We use the Feedforward-Pattern net with one hidden layer. Input vector (a compound from ACDn, ALn, Kn) contains 114 data samples from biometry measurements. Patients with three different monofocal intra-ocular lenses - BAUSCH & LOMB MI 60 (43 samples), CROMA ACR6D SE (55 samples), EYEOL UK LW 5752R (16 samples) were chosen.

4.1 Objective

To estimate whether the post-operative refraction based on ACD, AL and K values from preoperative biometry measurements will be larger or smaller than 0 diopters.

4.2 ANN training

Data were randomly divided into the three subsets (training set, validation set, testing set). ANN was trained on 80 samples, validated on 17 samples and tested on 17 samples causally chosen from InputVector with 114 samples.

ANN performance was calculated by cross-entropy which lot penalizes extremely inaccurate outputs and leads to good classifiers (Møller, 1993). Cross-entropy chart can be seen on Figure 5.

As a learning algorithm scaled conjugate gradient backpropagation was chosen. The data vector is fed to the input of ANN. After passing through the network the output of each neuron is calculated and the result is compared with the desired value. Mean squared error is calculated and previous layers and synaptic weights representing memory are corrected. This process repeats till minimum error between the real value and the desired value is reached. Scaled conjugate gradient backpropagation algorithm ensures rapid convergence of learning and using standard numerical optimization methods. More about this ANN learning algorithm can be found in (Tuckova, 2009; Pelusi, 2012).

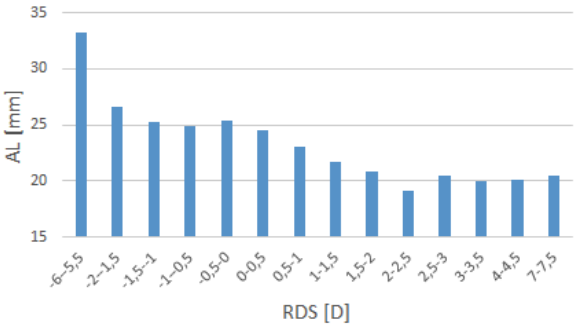


Figure 2. Relation of AL between RDS (Sramka, 2015).

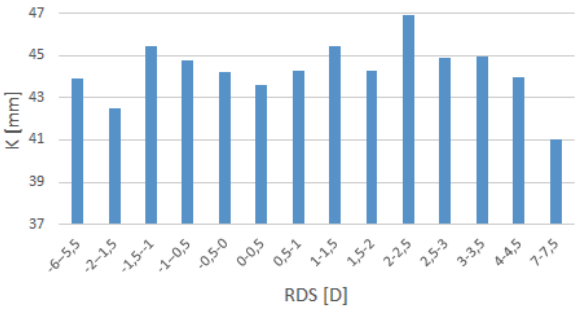


Figure 3. Relation of K between RDS (Sramka, 2015).

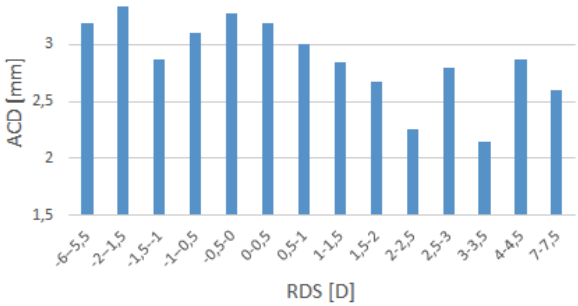


Figure 4. Relation of ACD between RDS (Sramka, 2015).

Algorithm for the best count of the hidden layer neurons was constructed and the ANN was tested for 1 to 50 neurons in the hidden layer. As can be seen in Figure 6 best performance was reached with 37 hidden layer neurons.

4.3 ANN settings

Following ANN settings which can also be seen in Figure 7 were used:

- Input neurons: 3
- Hidden layer neurons: 37
- Hidden layer transfer function: Hyperbolic tangent sigmoid
- Output layer neurons: 2
- Output layer transfer function: Soft max
- Number of training epochs: 16

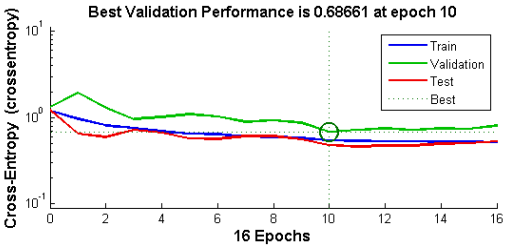


Figure 5. Relation between ANN performance and number of hidden layer neurons

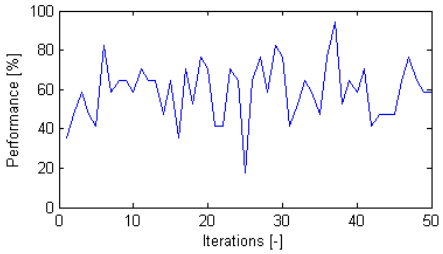


Figure 6. Relation between ANN performance and number of hidden layer neurons

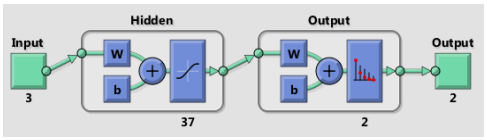


Figure 7. Scheme of Artificial Neural Network used in experiment

4.4 Results

On the Figure 8 there can be found that the overall accuracy of the testing is 94.1 %. Ten samples which represent 58.8 % of the testing set were correctly classified into Group 1 (Table 1). Six samples which represent 35.3 % of the testing set were correctly classified into Group 2 (Table 1). One sample which represents 5.9 % from Group 2 was incorrectly classified into Group 1.

5 Conclusions

This work deals with IOL power calculations during the cataract surgery. At present, intra-ocular lenses power is mostly being calculated by calculation formulas SRK/T, Holladay 1, Holladay 2, Haigis, Hoffer Q and others. All of these formulas using data measured by ultrasound or more often optical biometry and are able to provide acceptable results for the majority of the patients. This is based on fact that these formulas using constants which

Table 1. Group 1: Samples with Subjective Post-operative Refraction Larger than 0; Group 2: Samples with Subjective Post-operative Refraction Smaller than 0

Group 1	Group 2
SfS > 0	SfS < 0

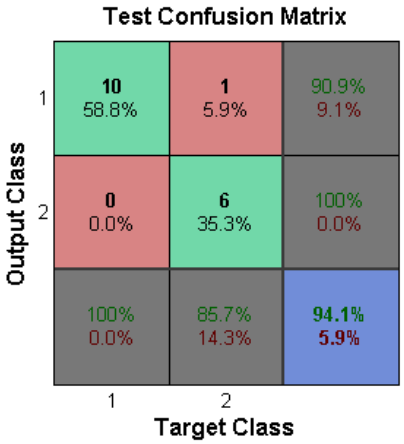


Figure 8. Test Confusion matrix. Green squares - samples correctly classified. Red squares - samples misclassified. Grey squares - Accuracy of each group. Blue square - overall accuracy (Sramka, 2015).

was derived by regression analysis. In the moment when any of input parameters has a value which is significantly unusual a problem can appear. In such a cases patient post-operative refraction can significantly deviate from intended target.

This work describes approach how to preoperatively compensate or indicate which samples of patients could be problematic in accurate intraocular lens calculations by classification of Artificial Neural Networks. Small and long eyes are used to test the ability of Artificial Neural Networks to classify input samples which are taken from pre-operative measurements to several groups which represent the probable post-operative result. In our experiment, Artificial Neural Network classifies samples into two groups. The first group is for data samples with a probable result in positive ranges of diopters and second group is for negative ranges. The accuracy of Artificial Neural Network, in this case, is 94.1 % and Artificial Neural Network seems like the instrument which has potential to improve an intra-ocular lens power calculation accuracy.

Based on the experiment Artificial Neural Networks seems to be the good solution for intra-ocular lens power compensation for non-standard eyes.

6 Future work

We will focus how to reach better accuracy in compensation inaccurate calculations by classification to more groups. Target is that each classification group would have an increment by 0.5 diopters, what is also a dioptric increment of IOL power usually given by manufacturers. Then could be inaccurate calculation compensate very exactly. Artificial Neural Network training could be tested for each IOL type and surgeon. Artificial Neural Network formula

selection could also be tested for the special cases. The special algorithm for compensation calculations for the patient with previous corneal refractive surgery could be designed.

Acknowledgment

The work was done in cooperation with Gemini Eye Clinic in Zlin. This article was supported by the Internal Grant Agency at TBU in Zlin, project No. IGA/CebiaTech/2016/007.

References

A. Abulafia, G. D. Barrett, M. Rotenberg, G. Kleinmann, A. Levy, O. Reitblat, D. D. Koch, L. Wang, and E. I. Assia. Intraocular lens power calculation for eyes with an axial length greater than 26.0 mm: Comparison of formulas and methods. *Journal of Cataract and Refractive Surgery*, 41(3):548 – 556, 2015. ISSN 0886-3350. doi:http://dx.doi.org/10.1016/j.jcrs.2014.06.033.

S. Bang, E. Edell, Q. Yu, K. Pratzer, and W. Stark. Accuracy of intraocular lens calculations using the {IOLMaster} in eyes with long axial length and a comparison of various formulas. *Ophthalmology*, 118(3):503 – 506, 2011. ISSN 0161-6420. doi:http://dx.doi.org/10.1016/j.ophtha.2010.07.008.

G. Carifi, F. Aiello, V Zygoura, N Kopsachilis, and M. Maurino. Accuracy of the refractive prediction determined by multiple currently available intraocular lens power calculation formulas in small eyes. *American Journal of Ophthalmology*, 159(3):577 – 583, 2015. ISSN 0002-9394. doi:http://dx.doi.org/10.1016/j.ajo.2014.11.036.

G.P. Clarke and J. Burmeister. Comparison of intraocular lens computations using a neural network versus the Holladay formula. *Journal of Cataract and Refractive Surgery*, 23(10): 1585–1589, 1997. cited By 6.

Y. Eom, S. Kang, J. S. Song, Y. Y. Kim, and H. M. Kim. Comparison of Hoffer Q and Haigis formulae for intraocular lens power calculation according to the anterior chamber depth in short eyes. *American Journal of Ophthalmology*, 157(4):818 – 824.e2, 2014. ISSN 0002-9394. doi:http://dx.doi.org/10.1016/j.ajo.2013.12.017.

B. N. Henderson, Pineda R., and S. H. Chen. *Essentials of cataract surgery*. SLACK Incorporated, Thorofare, New Jersey, second edition. edition, 2014. ISBN 978-1-63091-006-8.

P. Kuchynka. Oční lékařství. Grada, Praha, 1.vyd. edition, 2007. ISBN 80-247-1163-X.

M. F. Møller. Original contribution: A scaled conjugate gradient algorithm for fast supervised learning. *Neural Netw.*, 6(4):525 – 533, 1993.

Tivegna M. Pelusi, D. Optimal trading rules at hourly frequency in the foreign exchange markets. *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pages 341 – 348, 2012.

Roger Steinert. *Cataract surgery*. Saunders, Philadelphia, 3rd ed. edition, 2010. ISBN 978-141-6032-250.

M. Sramka. Artificial neural networks application in ophthalmology, intraocular lens power calculation improvement for patient undergoing cataract surgery. *Dissertation thesis statement*, 2015.

J. Tuckova. *Selected applications of the artificial neural networks at the signal processing.* Nakladatelství ČVUT, Praha, 2009. ISBN 978-80-01-04229-8.

J. Wang, Ch. Hu, and S. Chang. Intraocular lens power calculation using the {IOLMaster} and various formulas in eyes with long axial length. *Journal of Cataract and Refractive Surgery*, 34(2):262 – 267, 2008. ISSN 0886-3350. doi:http://dx.doi.org/10.1016/j.jcrs.2007.10.017.

L. Wang, M. Shirayama, X. J. Ma, T. Kohnen, and D. D. Koch. Optimizing intraocular lens power calculations in eyes with axial lengths above 25.0 mm. *Journal of Cataract and Refractive Surgery*, 37(11):2018 – 2027, 2011. ISSN 0886-3350. doi:http://dx.doi.org/10.1016/j.jcrs.2011.05.042.

Ch. Wong, L. Yuen, P. Tseng, and D. C. Y. Han. Outcomes of the Haigis-l formula for calculating intraocular lens power in Asian eyes after refractive surgery. *Journal of Cataract and Refractive Surgery*, 41(3):607 – 612, 2015. ISSN 0886-3350. doi:http://dx.doi.org/10.1016/j.jcrs.2014.06.034.

Tuning of Physiological Controller Motifs

Kristian Thorsen[1] Geir B. Risvoll[1] Daniel M. Tveit[1] Peter Ruoff[2] Tormod Drengstig[1]

[1]Department of Electrical Engineering and Computer Science, University of Stavanger, Norway,
{kristian.thorsen, geir.risvoll, daniel.m.tveit, tormod.drengstig}@uis.no
[2]Center for Organelle Research, University of Stavanger, Norway, peter.ruoff@uis.no

Abstract

Genetic manipulation is increasingly used to fine tune organisms like bacteria and yeast for production of chemical compounds such as biofuels and pharmaceuticals. The process of creating the optimal organism is difficult as manipulation may destroy adaptation and compensation mechanisms that have been tuned by evolution to keep the organisms fit. The continued progress in synthetic biology depends on our ability to understand, manipulate, and tune these mechanisms. Concepts from control theory and control engineering are very applicable to these challenges. From a control theoretic viewpoint, *disturbances rejection* and *set point tracking* describe how adaptation mechanisms relate to perturbations and to signaling events. In this paper we investigate a set regulatory mechanisms in the form of biochemical reaction schemes, so-called controller motifs. We show how parameters related to the molecular and kinetic mechanisms influence on the dynamical behavior of disturbance rejection and set point tracking of each controller motif. This gives insight into how a molecular controller motif can be tuned to a specified regulatory response.

Keywords: bioengineering, *biological systems*, *adaptation*

1 Introduction

1.1 Homeostasis, Disturbance Rejection and Set Point Tracking

Homeostasis is described as the mechanism behind the observed adaptation of an organism in a changing environment (Cannon, 1929; Langley, 1973). From a control theoretic point of view homeostasis can be described by the properties of disturbance rejection and set point tracking.

A physiological example of disturbance rejection is the intravenous/oral glucose tolerance test (IVGTT/OGTT), where the blood glucose concentration is measured at regular intervals after injecting/eating large amounts of glucose (Ackerman et al., 1964). If the blood glucose level is above a predefined level after a certain amount of time, the patient is often diagnosed as diabetic (Ame, 2014). Over the last half century, such disturbance rejection studies are reported in a vast number of publications, see e.g. (Larsen et al., 2003; Steele, 1959), and also a large number of mathematical models are made with the aim to capture the glucose and insulin dynamics, see e.g. the comprehensive review of (Ajmera et al., 2013). Both OGTT and IVGTT represent an impulse (or short time pulse) disturbance perturbation, whereas the chronic infusion of glucose (Topp et al., 2004) represent a stepwise disturbance. Another physiological example of adaptation to a stepwise perturbation change is the adaptation of light sensitivity of the eye, which includes both a compensatory change in pupillary size and an adaptation of the photochemical system in the rods and cones (Guyton and Hall, 2006).

Physiological examples where set point tracking is investigated are relatively rare, although set point determining mechanisms with respect to body temperature and metabolism have beed discussed (Briese, 1998; St Clair Gibson et al., 2005).

Regulatory mechanisms can today be synthetically modified or added to make organism better suitable for a specific job. Still, engineering of biochemical networks has not yet achieved the status and robustness as engineering of electrical and mechanical systems (Ang et al., 2010). From a synthetic biology perspective (Ang and McMillen, 2013; Ang et al., 2013), it is thus of vital importance to have insight into the biochemical mechanisms behind physiological regulatory systems. One possible way to gain such insight is to analyze both the disturbance rejection and set point tracking dynamics of such systems *in vivo*, as well as doing *in silico* studies based on different model candidates. The latter approach is a well known technique used in control engineering. We will in this paper start with the simplest form of biochemical networks with regulatory function and identify by model analysis and simulation how the dynamic response of such networks can be tuned.

1.2 Controller Motifs

A biochemical network with regulatory properties must in its simplest form include at least two components, i.e., state variables, one controlled component and a controller component. The controller component acts on the controlled component in a way that compensates for external disturbances. We have earlier presented a collection of simple two-component regulatory networks (Drengstig et al., 2012; Thorsen et al., 2013), and we have used the name *controller motifs* to describe them. These motifs consist of two chemical species, A and E,

both of them being formed and turned over. A may represent an intracellular compound which is subject to disturbances in the form of e.g. uncontrolled diffusive transport of A in and out of the cell, and E may represent a membrane bound compound such as a transporter protein as shown in Figure 1. Like many cellular compounds which is subject to strict regulation (due to e.g. toxicity if present in large amount), the concentration of A should not exceed or be less than some limits. By connecting the compounds A and E through cellular signaling events such as activation and inhibition, species A becomes the controlled variable, while species E becomes the manipulated variable.

Based on the direction of the E-mediated flow, the motifs fall into two categories termed *inflow* and *outflow* controllers. The complete set of possible inflow and outflow controller motifs are shown in Figure 2, and the *steady state* properties of these controllers were presented in (Drengstig et al., 2012). Based on the *type* of E-mediated inflow or outflow, the controllers are further divided into *activating* (inflow 1/3 and outflow 5/7) or *inhibiting* (inflow 2/4 and outflow 6/8) controller type, indicated by grey and white background in Figure 2, respectively.

In the following we will show how the parameters of the controller motifs, i.e. rate constants, Michaelis-Menten constants, activation constants and inhibition constants, influence on the dynamic performance, and show how it is possible to adjust the system's response similar to the tuning of industrial control systems.

2 Results

2.1 Dynamic Properties of Controller Motifs

The dynamic properties of a two component biochemical system (second order system) can be described in terms of the undamped natural frequency ω_n and the damping ratio ζ. To illustrate how these two parameters relate to the regulatory mechanisms in Figure 2, we use outflow controller 5 as an example. For unique identification, we

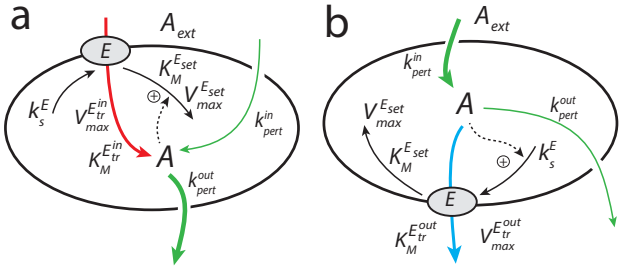


Figure 1. Illustration of a cell with a compound A being under homeostatic control by an inflow controller (panel a) or an outflow controller (panel b). Panel a: An inflow controller compensate for outflow perturbations, k_{pert}^{out} (thick green line), in A by adding more A through an E-mediated inflow (red line). Panel b: An outflow controller compensate for inflow perturbations, k_{pert}^{in} (thick green line), in A by removing excess of A through an E-mediated outflow (blue line).

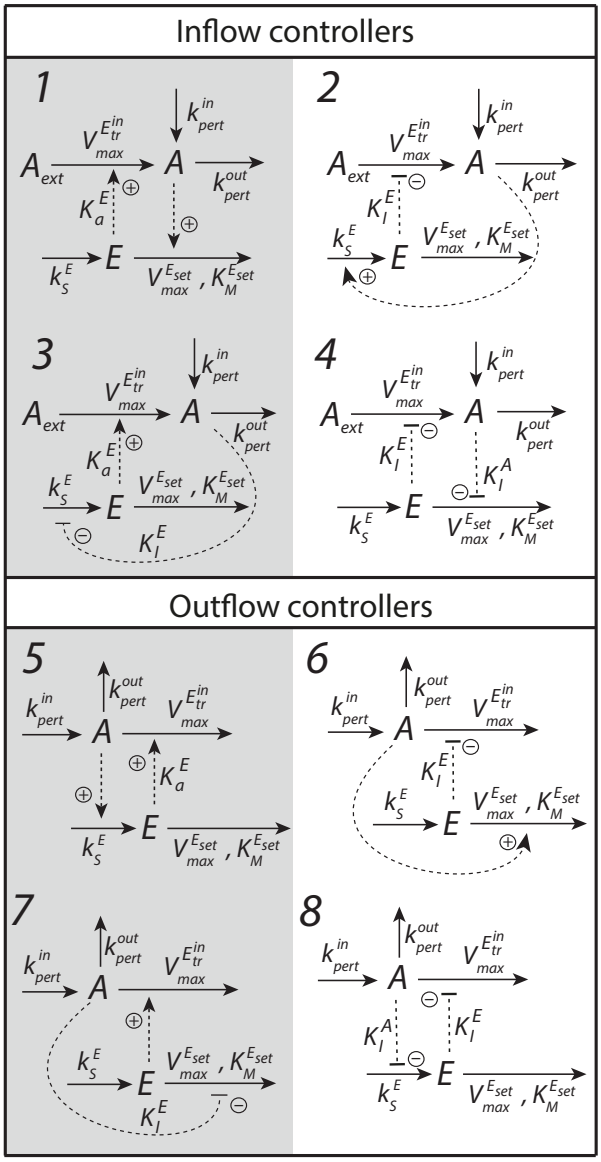


Figure 2. Set of two-component homeostatic controller motifs (Drengstig et al., 2012) classified as inflow and outflow controllers, where grey or white background indicate activating or inhibiting controller types, respectively.

apply subscript $_5$ on the appropriate parameters and variables, and hence, the nonlinear rate equations for an outflow controller 5 are given as (Drengstig et al., 2012):

$$\dot{A} = k_{pert}^{in} - k_{pert}^{out} \cdot A - V_{max}^{E_{tr},5} \cdot A \cdot \frac{E_5}{\left(K_a^{E_5} + E_5\right)} \tag{1}$$

$$\dot{E}_5 = k_s^{E_5} \cdot A - \frac{V_{max}^{E_{set},5} \cdot E_5}{\left(K_M^{E_{set},5} + E_5\right)} \tag{2}$$

As discussed in (Drengstig et al., 2012), the set point $A_{set}^{out,5}$ is found by assuming ideal (theoretical) conditions, i.e. $K_M^{E_{set},5}=0$ in (2), to give $A_{set}^{out,5}=\frac{V_{max}^{E_{set},5}}{k_s^{E_5}}$. Once the theoretical set point is established, we re-assume realistic conditions and reorganize (2) into the integral control law

equation $\dot{E}_5 = G_{i,5} \cdot (A_{set}^{out,5} - A_{meas})$. This allows us to identify the integral controller gain $G_{i,5}$ and the measurement signal A_{meas} as:

$$\dot{E}_5 = \underbrace{-k_s^{E_5} \cdot \frac{E_5}{K_M^{E_{set,5}} + E_5}}_{G_{i,5}} \cdot \left(\underbrace{\frac{V_{max}^{E_{set,5}}}{k_s^{E_5}}}_{A_{set}^{out,5}} - \underbrace{A \cdot \frac{K_M^{E_{set,5}} + E_5}{E_5}}_{A_{meas}} \right)$$

Note that the measurement signal A_{meas} actually includes information about the control signal E_5 which is not common in industrial control engineering. Note also that as long as $K_M^{E_{set,5}} > 0$, the actual value of A will be less than the theoretical set point $A_{set}^{out,5}$. Nevertheless, the set point tracking properties are good since the control error e, calculated as:

$$e = (A_{set}^{out,5} - A_{meas}) \tag{3}$$

is zero. The difference between the actual level of A and the theoretical set point A_{set} is termed *inaccuracy* (Thorsen, 2015). A general result valid for all controller motifs is that both rate constants for synthesis and degradation of E, i.e. k_s^E and $V_{max}^{E_{set}}$, are a part of the set point A_{set} (Drengstig et al., 2012). At the same time, one of these rate constants is also a part of the integral controller gain G_i.

In order to identify the parameters $\omega_{n,5}$ and ζ_5, we once again assume ideal conditions, i.e. $K_M^{E_{set,5}} = 0$, and continue by linearizing the model in (1) and (2) around an arbitrary working point A_{ss} and $E_{5,ss}$. Since the set point consist of two individual parameters, i.e. k_s^E and $V_{max}^{E_{set}}$, we select $V_{max}^{E_{set}}$ to be our input. We then find the closed looped transfer function from the Laplace transformed input $\Delta V_{max}^{E_{set,5}}(s)$ to the Laplace transformed output $\Delta A(s)$ as:

$$M(s) = \frac{\frac{\left(\left(k_{pert}^{out} + V_{max}^{E_{tr,5}} \right) \cdot V_{max}^{E_{set,5}} - k_{pert}^{in} \cdot k_s^{E_5} \right)^2}{V_{max}^{E_{set,5}} \cdot K_a^{E_5} \cdot V_{max}^{E_{tr,5}} \cdot k_s^{E_5}}}{s^2 + \frac{k_{pert}^{in} \cdot k_s^{E_5}}{V_{max}^{E_{set,5}}} \cdot s + \frac{\left(\left(k_{pert}^{out} + V_{max}^{E_{tr,5}} \right) \cdot V_{max}^{E_{set,5}} - k_{pert}^{in} \cdot k_s^{E_5} \right)^2}{V_{max}^{E_{set,5}} \cdot K_a^{E_5} \cdot V_{max}^{E_{tr,5}}}}$$

Using that $V_{max}^{E_{set,5}} = k_s^{E_5} \cdot A_{set}^{out,5}$, we find $\omega_{n,5}$ and ζ_5 as:

$$\omega_{n,5} = \frac{\sqrt{k_s^{E_5}} \cdot \left(\left(k_{pert}^{out} + V_{max}^{E_{tr,5}} \right) \cdot A_{set}^{out,5} - k_{pert}^{in} \right)}{\sqrt{K_a^{E_5} \cdot V_{max}^{E_{tr,5}} \cdot A_{set}^{out,5}}} \tag{4}$$

$$\zeta_5 = \frac{k_{pert}^{in} \sqrt{K_a^{E_5} \cdot V_{max}^{E_{tr,5}}}}{2 \cdot \sqrt{V_{max}^{E_{set,5}}} \cdot \left(\left(k_{pert}^{out} + V_{max}^{E_{tr,5}} \right) \cdot A_{set}^{out,5} - k_{pert}^{in} \right)} \tag{5}$$

From (4) and (5) we see that, depending on the perturbation levels (inflow versus outflow perturbations), it is possible to obtain negative values for $\omega_{n,5}$ and ζ_5. These negative values correspond to circumstances where the perturbation levels are such that the controller breaks down (Drengstig et al., 2012). Breakdown occurs when the net inflow perturbation is larger than the capacity of the outflow controller, i.e., greater than the maximum of the compensatory flow. In this case there is no stable equilibrium in the system and A integrates towards infinity. Such a state is unwanted and may very likely be toxic for the cell. In this case the values of $\omega_{n,5}$ and ζ_5 are invalid and have no physical meaning. Table 1 gives a summary of ω_n and ζ for the four inflow and four outflow controllers, together with the expression for each set point A_{set}.

Note that there is a close relationship between the expressions for ζ and ω_n for each controller, and thus, it is not possible to specify *both* ζ and ω_n independently.

Since controller 5 is an outflow controller, the inflow perturbation $\Delta k_{pert}^{in}(s)$ is considered the main disturbance, and the transfer function characterizing the disturbance rejection properties is:

$$N(s) = \frac{s}{s^2 + \frac{k_{pert}^{in} \cdot k_s^{E_5}}{V_{max}^{E_{set,5}}} \cdot s + \frac{\left(\left(k_{pert}^{out} + V_{max}^{E_{tr,5}} \right) \cdot V_{max}^{E_{set,5}} - k_{pert}^{in} \cdot k_s^{E_5} \right)^2}{V_{max}^{E_{set,5}} \cdot K_a^{E_5} \cdot V_{max}^{E_{tr,5}}}}$$

As expected, this transfer function has a zero in the origin, implicating homeostatic behavior and perfect adaptation (Drengstig et al., 2008).

2.2 Tuning of Individual Controllers

As shown in (Drengstig et al., 2012), the *steady state* performance of the individual controllers were found to be identical, given a certain set of parameter values. A related issue is to determine whether it is possible to tune the controllers to obtain identical *dynamical* performance using the theoretical design parameters in Table 1. Such tuning will be useful in synthetic biology. Also on a more fundamental level, if such tuning is possible it implies that it is impossible to infer the underlining network structure, i.e., the particular controller motif, responsible for an observed adaptive process by measuring the dynamical properties of the controlled variable alone.

We have selected to use the rate constants of the synthesis and degradation of the controller species, k_s^E and $V_{max}^{E_{set}}$, together with the rate constant of the E-mediated compensatory flow $V_{max}^{E_{tr}}$, as our tunable parameters. These parameters are relatively easy to tune from the perspective of synthetic biology and offer a greater tunable range than the parameters associated with the nonlinearities in the model (K_a^E, K_I^A, and K_I^E). To discuss one of the tunable parameters, the rate constant for synthesis of E, k_s^E, can in practice be modified by altering the promoter of the gene coding for E. One way to do this is a fixed tuning of the promoter itself, e.g. the *Cu*-dependent promoter of the CUP1-gene of *Saccharomyces Cerevisiae* can be modified by mutations to show wide range of different induction ratios (Thiele and Hamer, 1986). Another option is to use a dual mode promoter, a type of promoter who's regulation

Table 1. The set point A_{set}, natural undamped frequency ω_n and damping ratio ζ for controller motifs 1-8 in Figure 2 under theoretical conditions, i.e. $K_M^{E_{set}}=0$. For each controller we have added a subscript to the parameters for unique identification.

$A_{set}^{in,1}=\frac{k_s^{E_1}}{V_{max}^{E_{set,1}}}$	$\omega_{n,1}=\frac{\sqrt{V_{max}^{E_{set,1}}}\left(k_{pert}^{in}-k_{pert}^{out}A_{set}^{in,1}+V_{max}^{E_{tr,1}}A_{ext}\right)}{\sqrt{K_a^{E_1}V_{max}^{E_{tr,1}}A_{ext}}}$	$\zeta_1=\frac{k_{pert}^{out}\sqrt{K_a^{E_1}V_{max}^{E_{tr,1}}A_{ext}}}{2\sqrt{V_{max}^{E_{set,1}}}\left(k_{pert}^{in}-k_{pert}^{out}A_{set}^{in,1}+V_{max}^{E_{tr,1}}A_{ext}\right)}$
$A_{set}^{in,2}=\frac{V_{max}^{E_{set,2}}}{k_s^{E_2}}$	$\omega_{n,2}=\frac{\sqrt{k_s^{E_2}}\left(k_{pert}^{out}A_{set}^{in,2}-k_{pert}^{in}\right)}{\sqrt{K_I^{E_2}V_{max}^{E_{tr,2}}A_{ext}}}$	$\zeta_2=\frac{k_{pert}^{out}\sqrt{K_I^{E_2}V_{max}^{E_{tr,2}}A_{ext}}}{2\sqrt{k_s^{E_2}}\left(k_{pert}^{out}A_{set}^{in,2}-k_{pert}^{in}\right)}$
$A_{set}^{in,3}=\frac{k_s^{E_3}K_I^A}{V_{max}^{E_{set,3}}}-K_I^A$	$\omega_{n,3}=\frac{\sqrt{V_{max}^{E_{set,3}}}\left(k_{pert}^{in}-k_{pert}^{out}A_{set}^{in,3}+V_{max}^{E_{tr,3}}A_{ext}\right)}{\sqrt{\left(K_I^A+A_{set}^{in,3}\right)V_{max}^{E_{tr,3}}K_a^{E_3}A_{ext}}}$	$\zeta_4=\frac{k_{pert}^{out}\sqrt{\left(K_I^A+A_{set}^{in,4}\right)V_{max}^{E_{tr,4}}K_I^{E_4}A_{ext}}}{2\sqrt{k_s^{E_4}}\left(k_{pert}^{out}A_{set}^{in,4}-k_{pert}^{in}\right)}$
$A_{set}^{in,4}=\frac{V_{max}^{E_{set,4}}K_I^A}{k_s^{E_4}}-K_I^A$	$\omega_{n,4}=\frac{\sqrt{k_s^{E_4}}\left(k_{pert}^{out}A_{set}^{in,4}-k_{pert}^{in}\right)}{\sqrt{\left(K_I^A+A_{set}^{in,4}\right)V_{max}^{E_{tr,4}}K_I^{E_4}A_{ext}}}$	$\zeta_3=\frac{k_{pert}^{out}\sqrt{\left(K_I^A+A_{set}^{in,3}\right)V_{max}^{E_{tr,3}}K_a^{E_3}A_{ext}}}{2\sqrt{V_{max}^{E_{set,3}}}\left(k_{pert}^{in}-k_{pert}^{out}A_{set}^{in,3}+V_{max}^{E_{tr,3}}A_{ext}\right)}$
$A_{set}^{out,5}=\frac{V_{max}^{E_{set,5}}}{k_s^{E_5}}$	$\omega_{n,5}=\frac{\sqrt{k_s^{E_5}}\left(\left(k_{pert}^{out}+V_{max}^{E_{tr,5}}\right)A_{set}^{out,5}-k_{pert}^{in}\right)}{\sqrt{K_a^{E_5}V_{max}^{E_{tr,5}}A_{set}^{out,5}}}$	$\zeta_5=\frac{k_{pert}^{in}\sqrt{K_a^{E_5}V_{max}^{E_{tr,5}}}}{2\sqrt{V_{max}^{E_{set,5}}}\left(\left(k_{pert}^{out}+V_{max}^{E_{tr,5}}\right)A_{set}^{out,5}-k_{pert}^{in}\right)}$
$A_{set}^{out,6}=\frac{k_s^{E_6}}{V_{max}^{E_{set,6}}}$	$\omega_{n,6}=\frac{\sqrt{V_{max}^{E_{set,6}}}\left(k_{pert}^{in}-k_{pert}^{out}A_{set}^{out,6}\right)}{\sqrt{K_I^{E_6}V_{max}^{E_{tr,6}}A_{set}^{out,6}}}$	$\zeta_6=\frac{k_{pert}^{in}\sqrt{K_I^{E_6}V_{max}^{E_{tr,6}}}}{2\sqrt{k_s^{E_6}}\left(k_{pert}^{in}-k_{pert}^{out}A_{set}^{out,6}\right)}$
$A_{set}^{out,7}=\frac{V_{max}^{E_{set,7}}K_I^A}{k_s^{E_7}}-K_I^A$	$\omega_{n,7}=\frac{\sqrt{k_s^{E_7}}\left(\left(k_{pert}^{out}+V_{max}^{E_{tr,7}}\right)A_{set}^{out,7}-k_{pert}^{in}\right)}{\sqrt{\left(A_{set}^{out,7}+K_I^A\right)V_{max}^{E_{tr,7}}K_a^{E_7}A_{set}^{out,7}}}$	$\zeta_7=\frac{k_{pert}^{in}\sqrt{\left(A_{set}^{out,7}+K_I^A\right)V_{max}^{E_{tr,7}}K_a^{E_7}}}{2\sqrt{k_s^{E_7}A_{set}^{out,7}}\left(\left(k_{pert}^{out}+V_{max}^{E_{tr,7}}\right)A_{set}^{out,7}-k_{pert}^{in}\right)}$
$A_{set}^{out,8}=\frac{k_s^{E_8}K_I^A}{V_{max}^{E_{set,8}}}-K_I^A$	$\omega_{n,8}=\frac{\sqrt{V_{max}^{E_{set,8}}}\left(k_{pert}^{in}-k_{pert}^{out}A_{set}^{out,8}\right)}{\sqrt{\left(A_{set}^{out,8}+K_I^A\right)V_{max}^{E_{tr,8}}K_I^{E_8}A_{set}^{out,8}}}$	$\zeta_8=\frac{k_{pert}^{in}\sqrt{\left(A_{set}^{out,8}+K_I^A\right)V_{max}^{E_{tr,8}}K_I^{E_8}}}{2\sqrt{V_{max}^{E_{set,8}}A_{set}^{out,8}}\left(k_{pert}^{in}-k_{pert}^{out}A_{set}^{out,8}\right)}$

of protein production depends on two activators. One activator would be the control variable A and another would be a chemical compound that can be meticulously added to the growth medium to achieve a certain level of gene transcription and production of E, represented in the model as the value of k_s^E. One such promoter controlled by Testosterone and IPTG (isopropyl β-D-1-thiogalactopyranoside) has recently been developed (Mazumder and McMillen, 2014).

In order to best tune the parameters we have to know about the operational limits of the system. For this purpose, we define as in (Drengstig et al., 2012) an upper limit for the maximum compensatory flux, $j_{A,max}=10$, corresponding to a maximum level of $E_{max}=15$ for the activating controllers 1, 3, 5 and 7, and corresponding to $E_{min}=0$ for the inhibiting controllers 2, 4, 6 and 8. We assume further that the set point of A is $A_{set}=1.0$, the external concentration is $A_{ext}=2$. The kinetic constants for activation and inhibition are chosen to avoid saturation effects: $K_a^E=2$, $K_I^A=0.1$ and $K_I^E=1.0$. Moreover, the working point of perturbations is specified as $k_{pert}^{in}=2/k_{pert}^{out}=5$ for inflow controllers and $k_{pert}^{in}=5/k_{pert}^{out}=2$ for outflow controllers. Given these overall system parameters, the tuning procedure of each individual controller motif is based on specifying ζ (or ω_n, but not both) in a similar way as the pole placement method, and determine the last three parameter values of each motif, i.e. $V_{max}^{E_{tr}}$, $V_{max}^{E_{set}}$ and k_s^E.

To illustrate, we specify *two* different dynamical responses in the concentration of A for a step in A_{set}, i.e. one critically damped ($\zeta=1$) and one underdamped ($\zeta=0.2$ corresponding to 50% overshoot) response. A strongly underdamped system overshoots when adapting a change in set point, but shows considerably better disturbance rejection than a critically damped system. Thus, tuning for the latter may be of interest in many biological systems.

We illustrate the procedure in detail by continuing on the outflow controller 5 example, and start by considering the rate expression for the compensatory flux, j_A, from (1):

$$j_A=V_{max}^{E_{tr,5}}\cdot A\cdot\frac{E_5}{\left(K_a^{E_5}+E_5\right)} \tag{6}$$

By setting $j_A=j_{A,max}=10$ and inserting $E_5=E_{5,max}=15$, $A=A_{set}^{out,5}=1$ and $K_a^{E_5}=2$ into (6), gives $V_{max}^{E_{tr,5}}=11.33$. Using the mathematical expressions for $A_{set}^{out,5}$ and ζ_5 tabulated in Table 1, we find $V_{max}^{E_{set,5}}=2.04$ and $k_s^{E_5}=2.04$ for

ζ_5=1 and $V_{max}^{E_{set},5}$=51.0 and $k_s^{E_5}$=51.0 for ζ_5=0.2, see Table 2.

Table 2. The parameters $V_{max}^{E_{tr}}$, $V_{max}^{E_{set}}$, k_s^E and the integral controller gain G_i (in grey) for each controller motif specified for critical damped response ζ=1 and underdamped response ζ=0.2. The other parameters are defined in the main text.

		$V_{max}^{E_{tr}}$	$V_{max}^{E_{set}}$	k_s^E	G_i
Critically damped, ζ=1	Inflow 1	5.67	2.04	2.04	2.04
	Inflow 2	5.00	6.94	6.94	-6.94
	Inflow 3	5.67	2.24	24.68	2.04
	Inflow 4	5.00	84.03	7.64	-6.94
Under-damped, ζ=0.2	Inflow 1	5.67	51.0	51.0	51.0
	Inflow 2	5.00	173.6	173.6	-173.6
	Inflow 3	5.67	56.1	617.1	51.0
	Inflow 4	5.00	2100.7	191.0	-173.6
Critically damped, ζ=1	Outflow 5	11.33	2.04	2.04	-2.04
	Outflow 6	10.00	6.94	6.94	6.94
	Outflow 7	11.33	24.68	2.24	-2.04
	Outflow 8	10.00	7.64	84.03	6.94
Under-damped, ζ=0.2	Outflow 5	11.33	51.0	51.0	-51.0
	Outflow 6	10.00	173.6	173.6	173.6
	Outflow 7	11.33	617.1	56.1	-51.0
	Outflow 8	10.00	191.0	2100.7	173.6

This corresponds to an integral controller gain of $G_{i,5}$=−2.04 and $G_{i,5}$=−51.0, respectively, and a response time of T_r≈0.8 seconds ($\omega_{n,5}$=2.5) and T_r≈0.1 seconds ($\omega_{n,5}$=12.5). The simulation results shown as black curves in panels **c**, and **d** in Figure 3, verify the tuning specifications, both with respect to overshoot and response time.

In order to compare the individual performance of each controller, the above described tuning specifications are applied for all controllers, and the results are shown in Table 2 and verified by simulation in Figure 3.

Note the identical values for G_i (greyed out in Table 2) for all the activating (inflow 1/3 and outflow 5/7) and all the inhibiting (inflow 2/4 and outflow 6/8) controllers, respectively. Note also the opposite signs for activating and inhibiting inflow and outflow controllers, respectively, which is due to the combination of controller type (*activating/inhibiting*) and controller configuration (*inflow/outflow*).

The responses in Figure 3 clusters into two groups, where the first group is the E-activating inflow controllers 1/3 (black and red curves in Figures 3**a** and 3**b**) and the E-inhibiting outflow controllers 6/8 (blue and green curves in Figures 3**c** and 3**d**). The second group is the E-inhibiting inflow controllers 2/4 (blue and green curves in Figures 3**a** and 3**b**) and the E-activating outflow controllers 5/7 (black and red curves in Figures 3**c** and 3**d**). The reason why equally tuned controllers behaves slightly different is due to the nonlinearity of each individual controller combined with a relative large set point step change.

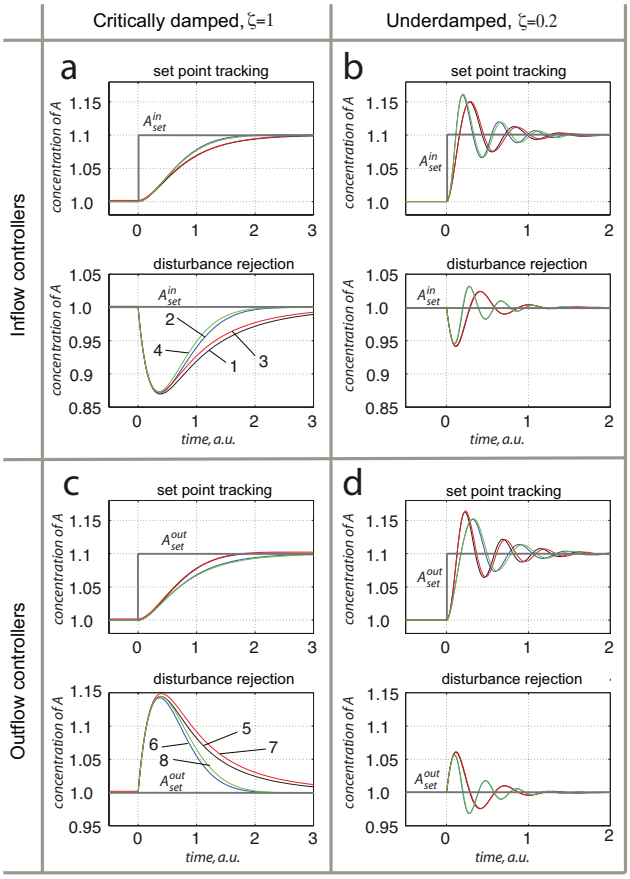


Figure 3. Dynamic properties of inflow and outflow controllers showing the response in concentration of species A. The color codes for the different inflow controller are: 1=black, 2=blue, 3=red and 4=green, and the color codes for the different outflow controllers are: 5=black, 6=blue, 7=red and 8=green. For the *set point tracking* curves, the set point changes from A_{set}=1.0 to A_{set}=1.1 at t=0. For the *disturbance rejection* curves, the disturbance is a unit step change from 5 to 6 at t=0 in k_{pert}^{out} for inflow controllers and in k_{pert}^{in} for outflow controllers. Panels **a** and **b**: Set point tracking (upper) and disturbance rejection (lower) responses for inflow controllers tuned for critically damped (ζ=1) and underdamped (ζ=0.2) responses, using the parameters shown in Table 2. Panels **c** and **d**: Set point tracking (upper) and disturbance rejection (lower) responses for outflow controllers tuned for critically damped (ζ=1) and underdamped (ζ=0.2) responses, using the parameters shown in Table 2.

From Table 1 we see that the inflow and outflow perturbations come into the expressions of ω_n and ζ in different ways. To visualize the effect of varying level of perturbation, Figure 4 shows dynamic responses of inflow controller 3 for k_{pert}^{out}={3,5,7} (Figure 4**a**) and outflow controller 6 for k_{pert}^{in}={5,7,9} (Figure 4**b**). The effect of increased k_{pert}^{out} for inflow controller 3 is slower dynamics with less damped response. On the other hand, outflow controller 6 shows faster dynamics together with more underdamped response at increased k_{pert}^{in} levels.

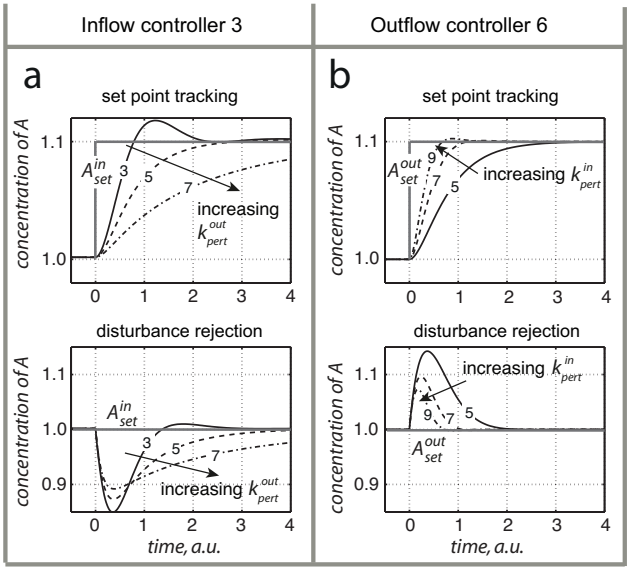


Figure 4. Set point tracking (upper) and disturbance rejection dynamics (lower) of species A using inflow controller 3 (panel **a**) and outflow controller 6 (panel **b**) at different level of outflow and inflow perturbations, respectively. The set point change is a step from A_{set}=1.0 to A_{set}=1.1 at t=0 and the disturbance is a step increase of 1 from original value at t=0. In panel **a** the labeling on the curves corresponds to outflow perturbations of $k^{out}_{pert} \in \{3,5,7\}$. In panel **b** the labeling on the curves corresponds to inflow perturbation of $k^{in}_{pert} \in \{5,7,9\}$.

3 Conclusions

We have shown how a set of homeostatic controller motifs can be tuned, in a similar way as in industrial control systems, to exhibit a specified dynamic response with respect to overshoot δ and response time T_r. We have also shown analytically and through simulations how *i*) the level of inflow/outflow disturbances and *ii*) the values of different rate constants influence on the set point tracking properties. The corresponding disturbance rejection properties is also studied through simulations using a unit step input signal in the disturbance.

An important implication of the fact that all controller motifs can show identical dynamic responses is that one cannot postulate a specific controller motif based on measurement of disturbance rejection and/or set point tracking alone. The motif type, i.e. inflow or outflow, activating or inhibiting, rest on how the molecular mechanisms behind the controller interact and not on the system's ability to show a specific response. The specific response of physiological regulatory system is a result of tuning the system's kinetic parameters and the strength of the perturbation.

There is a great effort going on in both academia and industry to genetically manipulate organisms to produce useful bioproducts. One of the landmark studies published in Science last year was the implementation of the complete biosynthesis of opioids in yeast (Galanie et al., 2015; Service, 2015). Opioids like morphine are the primary drugs used for treatment of severe pain and pain management, and production depends on the cultivation of opium poppies. While the implementation of opioid biosynthesis in yeast is a tremendous achievement, it still requires an improvement in overall yield by a factor of $7 \cdot 10^6$ to compete with poppies (Galanie et al., 2015). Great improvements are expected (Galanie et al., 2015), but this will require an intricate tuning of the different parts of the biosynthesis pathway.

From a synthetic biology point of view, the work in this paper creates a basis one can use to identify which and how properties of a reaction and participating proteins/enzymes contributes to the dynamical response. For instance, the natural undamped frequency ω_n, which is important for the swiftness of a controller motif, will for outflow controller 5 increase if we by some means manage to increase the production of E (increase k^E_s) by e.g. increasing the expression of mRNA coding for E (as shown in Table 1, a change in k^E_s will also change the set point). A related example of such is reported in (Ang et al., 2010), where a two promotor network system is constructed *in silico* from realizable parts within the bacterium *Escherichia coli*. The network includes both basal rates and activated/repressed regulatory inputs, and hence, the network share similarities with inflow controller 2 in Figure 2. Two requirements were used as tuning criteria for the network, i.e. ζ=1 (critically damped) and large ω_n indicating a response time T_r as short as possible. In order to obtain the necessary approximate zero order degradation of the repressor R (corresponding to our species E), two effectors I_1 and I_2 are included in order to force the repressor to work at saturated conditions, i.e. corresponding to the theoretical conditions, $K^{E_{set}}_M$=0, used in this paper.

An alternative approach to tuning is given in (Ang et al., 2013), where the tuning is related to the so-called *response curves*. These are steady state relationships between an input and an output variable, e.g. the molecular concentration of a transcription factor protein and the expressed protein, respectively, and not time dependent tuning as discussed in this paper. However, variations in kinetic parameter values results in different steady state relationships.

References

E. Ackerman, J.W. Rosevear, and W.F. McGuckin. A Mathematical Model of the Glucose-tolerance test. *Physics in Medicine and Biology*, 9(2):203–213, 1964.

I. Ajmera, M. Swat, C. Laibe, N. Le Novère, and V. Chelliah. The impact of mathematical modeling on the understanding of diabetes and related complications. *CPT: Pharmacometrics & Systems Pharmacology*, 2 (7):14, 2013.

Standards of Medical Care in Diabetes–2014. American Diabetes Association, 2014. *Diabetes Care* 37:S14–S80.

J. Ang and D.R. McMillen. Physical Constraints on Biological Integral Control Design for Homeostasis and

Sensory Adaptation. *Biophysical Journal*, 104(2):505–515, 2013.

J. Ang, S. Bagh, B.P. Ingalls, and D.R. McMillen. Considerations for using integral feedback control to construct a perfectly adapting synthetic gene network. *Journal of Theoretical Biology*, 266(4):723–738, 2010.

J. Ang, E. Harris, B.J. Hussey, R. Kil, and D.R. McMillen. Tuning Response Curves for Synthetic Biology. *ACS Synthetic Biology*, 2(10):547–567, 2013.

E. Briese. Normal body temperature of rats: the setpoint controversy. *Neuroscience & Biobehavioral Reviews*, 22(3):427–436, 1998.

W.B. Cannon. Organization for physiological homeostasis. *Physiological reviews*, IX:399–431, 1929.

T. Drengstig, I.W. Jolma, X.Y. Ni, K. Thorsen, X.M. Xu, and P. Ruoff. A Basic Set of Homeostatic Controller Motifs. *Biophysical Journal*, 103:2000–2010, 2012.

T. Drengstig, H.R. Ueda, and P. Ruoff. Predicting Perfect Adaptation Motifs in Reaction Kinetic Networks. *Journal of Physical Chemistry B*, 112(51):16752–16758, 2008.

S. Galanie, K. Thodey, I.J. Trenchard, M.F. Interrante, and C.D. Smolke. Complete biosynthesis of opioids in yeast. *Science*, 349(6252):1095–1100, 2015.

A.C. Guyton and J.E. Hall. *Textbook of Medical Physiology*. Elsevier Saunders, Philadelphia, PA, USA, 11 edition, 2006.

L.L. Langley. *Homeostasis: Origins of the concept*. John Wiley & Sons, 1973.

M.O. Larsen, B. Rolin, M. Wilken, R.D. Carr, and C.F. Gotfredsen. Measurements of insulin secretory capacity and glucose tolerance to predict pancreatic β-cell mass in vivo in the nicotinamide/streptozotocin Göttingen minipig, a model of moderate insulin deficiency and diabetes. *Diabetes*, 52:118–123, 2003.

M. Mazumder and D.R. McMillen. Design and characterization of a dual-mode promoter with activation and repression capability for tuning gene expression in yeast. *Nucleic Acids Research*, 42(14):9514–9522, 2014.

R.F. Service. Modified yeast produce opiates from sugar. *Science*, 349(6249):677–677, 2015.

A. St Clair Gibson, J.H. Goedecke, Y.X. Harley, L.J. Myers, M.I. Lambert, T.D. Noakes, and E.V. Lambert. Metabolic setpoint control mechanisms in different physiological systems at rest and during exercise. *Journal of Theoretical Biology*, 236:60–72, 2005.

R. Steele. Influences of glucose loading and of injected insulin on hepatic glucose output. *Annals of the New York Academy of Sciences*, 82:420–430, 1959.

D.J. Thiele and D.H. Hamer. Tandemly duplicated upstream control sequences mediate copper-induced transcription of the Saccharomyces cerevisiae copper-metallothionein gene. *Molecular and Cellular Biology*, 6(4):1158–1163, 1986.

K. Thorsen. *Controller Motifs for Homeostatic Regulation and Their Applications in Biological Systems*. PhD thesis, University of Stavanger, 2015.

K. Thorsen, P. Ruoff, and T. Drengstig. Control theoretic properties of physiological controller motifs. In *2013 International Conference on System Science and Engineering (ICSSE)*, pages 165–170. IEEE, 2013.

B.G. Topp, M.D. McArthur, and D.T. Finegood. Metabolic adaptations to chronic glucose infusion in rats. *Diabetologia*, 47(9):1602–1610, 2004.

How does Modern Process Automation understand the Principles of Microbiology and Nature

Ari Jääskeläinen[1] Risto Rissanen[1] Asmo Jakorinne[1] Anssi Suhonen[1] Tero Kuhmonen[1]
Tero Reijonen[1] Eero Antikainen[1] Anneli Heitto[2] Elias Hakalehto[2]

[1]School of Technology, Savonia University of Applied Sciences, Finland, ari.jaaskelainen@savonia.fi
[2]Finnoflag Oy, Finland, elias.hakalehto@gmail.com

Abstract

Future biorefineries will work according to the principles of Nature, using microbes and enzymes for valorizing wastes and other biomass into biofuels, other bioenergy substances, organic platform chemicals, and organic fertilizers. A mobile biorefinery pilot plant was engineered and manufactured in Finland and tested in Finland, Poland and Sweden with various biowastes within ABOWE project. The main purpose of the ABOWE Biorefinery pilot plant tests was to give a reliable "proof of concept" on the industrially important substances producible in a sustainable way. This goal was achieved successfully, and several overall difficulties were overcome during the testing in three countries.

Keywords: bioprocess, biorefinery, biowaste, consolidated bioprocessing, undefined mixed cultures, anaerobiosis

1 Introduction

Biotechnology deals with processes that are based on naturally occurring phenomena. Bastin and Dochain (1990) define bioreactor as a tank in which several biological reactions occur simultaneously in a liquid medium. The biological reactions which are involved in the process can be classified as microbial growth reactions/microbiological reactions and enzyme catalyzed reactions/biochemical reactions. The growth of the microorganisms (bacteria, yeasts, etc.) proceeds by consumption of appropriate nutrients or substrates such as carbon, nitrogen and oxygen. Preferably environmental conditions (temperature, pH, etc.) are optimized. The cell growth is associated with the enzymatic reactions in which some reactants are transformed into products through the catalytic action of enzymes. (Bastin and Dochain, 1990)

Favorable environmental conditions for microorganisms should be maintained all the time. Exceeding boundary values regarding conditions leads to slowing of the production rate, and there is often no other way than to readjust or restart. The failure of the process does not always require even crossing the threshold, but this may also happen due to circumstances in which some of the harmful microbes take over.

To process control and automation the multiple variables in a mixed microbial culture set high demands. Parameters in terms of automation are, *e.g.*, pH, temperature, sugar balance, dissolved oxygen and gas mixtures. Each of the above-mentioned variables has its own control and adjustment aspects. This is the so-called multivariable system, in which there are dependencies between parameters. Each parameter's effects on microbial activity should be understood in order to get the optimal outcome from the automation and process design. The principles of fuzzy logic might be needed. Temperature control alone is already a challenging thing, because the process itself generates heat, which should be predicted. Not to mention the other variables. The phenomena are very unstable and cross-effects are common. For control engineering and process control these are demanding cases to adjust and measure the variables in order to reach the desired outcomes.

Processes cannot be adjusted only with traditional methods used in heavy industry but events have to be controlled with a technique called bioautomation. In addition to the physical sensors there is information required that might not even be possible to get with concentration analysis measurements. There is a need to go deeper into the events even at the cellular level for which micro- and nanotechnology bring along new opportunities.

Knowledge of processes is very important in automation engineering. Microbiology often deals with more difficult phenomena than regular process chemistry does. The problems are almost impossible to be solved otherwise than through close cooperation between experts from various fields.

Future biorefineries will work according to the principles of Nature, using microbes and enzymes for valorizing wastes and other biomass into biofuels, other bioenergy substances, organic platform chemicals, and organic fertilizers. The concept of 'waste' will become unnecessary in industries, communities, agriculture and forestry as all materials are being refined and recycled. The ABOWE (Implementing Advanced Concepts for

Biological Utilization of Waste, 12/2012-12/2014) project, led by Savonia University of Applied Sciences, and its two pilot plants, have paved the way for this industrial revolution. ABOWE project belonged to the EU Baltic Sea Region Programme 2007-2013. (Jääskeläinen and Hakalehto, 2015)

The novel biorefinery concept, innovated and developed by Adjunct Professor Elias Hakalehto, (Finnoflag Oy and University of Eastern Finland), was one of the two platforms of the ABOWE project. The second platform was biogas production with the dry digestion technology piloted under the supervision of Ostfalia University of Applied Sciences, Germany. ABOWE was an extension project for REMOWE project (Regional Mobilization of Sustainable Waste to Energy Production 9/2009-12/2012) to continue with these two promising technologies to piloting phase. (Hakalehto *et al*, 2016a)

Savonia University of Applied Sciences invested in the ABOWE project in a mobile biorefinery pilot plant. In this biorefinery plant biodegradable wastes are valorized to valuable chemical and energy products with the aid of microbes and their enzymes, in the same way as in the natural environments.

The pilot plant engineering team consisted of Finnoflag Oy experts and Savonia's engineering teachers, project engineers and engineering students. The project team was built so that teaching was integrated as widely as possible already at the planning stage of the process. Versatile knowledge of process and instrumentation, layout, mechanical, electrical, automation, IT, environmental and manufacturing was combined for the pilot plant engineering and manufacturing during 2013.

Savonia manufactured the pilot plant in front of its educational workshop. In the pilot plant manufacturing, several locally operating industrial enterprises were participating as component suppliers. Also many trainees from Savo Vocational College, Finland, participated in the manufacturing of the pilot plant. Altogether over 50 persons were involved in the pilot plant engineering and manufacturing in Finland. There was not much available model, and practical experience about effects of various functions was mostly lacking. Savonia's personnel instructed the project, brought engineering know how to the project and performed equipment purchases. (Jääskeläinen and Hakalehto, 2018)

The biorefinery pilot plant was completed in January 2014 and was tested in Finland with waste water treatment sludge at a fluting factory, in Poland with potato waste from chips factory and in Sweden with chicken manure and slaughterhouse waste during 2014.

2 Materials and Methods

Objectives of the ABOWE project regarding the mobile biorefinery pilot plant were to test

- Effective pretreatments and hydrolysis of various biodegradable wastes.
- Enhanced natural microbial bioprocess for the upstream production of fuels and chemicals.
- Preliminary planning of the simultaneous product collection.

The goal of the ABOWE project and the mobile pilot plant was to provide "proof of concept" on the ways, how biomass waste materials could be used as raw materials. The products were biofuels, bioenergy, organic platform chemicals and organic fertilizers. These should be produced in an economically feasible way, with the help of micro-organisms. (Hakalehto *et al*, 2016b)

The biorefinery process is novel in terms of improved productivity, low initial investment costs and versatile product opportunities. The production exploits results from the research conducted in the Finnoflag laboratory since 1997 with the PMEU enhanced cultivation unit (Portable Microbe Enrichment Unit), and in larger vessels. As various products are produced faster, the production plant size reduces enabling lower investment. Moreover, the total duration of the process can be shortened and end product concentrations increased. (Hakalehto *et al*, 2016b)

There are four main tanks in the mobile biorefinery pilot plant's process (Hakalehto *et al*, 2016b):

1. HOMOGENIZER is the first main tank. Attached to it is a biomass crusher and the tank has an effective mixer. Water content can be adjusted partially in Homogenizer from a water tap. Homogenizer is also one of the three recycled and modified tanks in the pilot plant used for the upstream bioprocess. In Homogenizer various biomasses are mechanically broken in micro- and macroscale. Their dry weight is measured in the on-site lab room and the total masses of solid and liquid substrates are measured with a weighing sensor installed in the leg of the tank.

2. HYDROLYZER is a reactor with thermostat and pH control for producing, maintaining and adjusting the optimal conditions for chemical and/or enzymatic hydrolysis of the macromolecules in the substrate biomasses. Main operating parameters are the water content, fill in level, temperature (can be lifted up to 80 degrees Celsius), pH of the biomass, viscosity and the hydrolysis time. Hydrolyzer temperature is controlled with a control sequence whose parameters can be set from the control room. The purpose of the sequence is to perform the hydrolysis of the raw material and to kill harmful microbes so that the batch would be ready to be fed in the Bioreactor. Also pH is started to be adjusted for the coming process phases.

3. BIOREACTOR is the sole entirely novel big tank in the pilot plant. It has been manufactured by (Brandente Oy, Kuopio, Finland) according to the instructions of the innovator (Finnoflag Oy, Siilinjärvi, Finland) and Savonia. The patented design is based on numerous bioprocess runs in Finnoflag Oy's laboratory projects preceding the ABOWE project. Different homogenized and hydrolyzed biomasses are processed in adjustable gas conditions in the Bioreactor in order to produce biofuels, biogases and biochemicals by the metabolic activities of bacteria and other micro-organisms. The volume of the tank is around 300 liters of which effective volume is around 200 liters.

An optimal gas mixture is introduced to the Bioreactor from gas bottles and air compressor via a gas mixer. Typical gases besides the air are carbon dioxide and nitrogen depending on the specific requirements of the microbial culture. An optimal mixture is introduced via aeration discs and rings from two layers in the Bioreactor. The gas flow is at the same time performing the airlift principle in the Bioreactor mixing. There is not any mechanical stirring system, *e.g.*, to avoid shear forces in the process broth.

There is a two-step process for producing the microbiological inocula: first in the PMEU equipment (Portable Microbe Enrichment Unit) (Samplion Oy, Siilinjärvi, Finland), and then in the seed fermenters connected to the Bioreactor. In PMEU it is possible to get homogenous cultures into same active growth phase in a few hours of cultivation (Hakalehto and Heitto, 2012). Besides ABOWE version, the PMEU has been productized earlier for environmental, clinical, food, and other hygienic control purposes (Hakalehto, 2012).

Additional substances and pH adjustment chemicals are fed to a circulation from the Bioreactor back to the Bioreactor, run by a circulation pump. Thus substances can easier be mixed with the mass and local higher concentrations are not occurring in the Bioreactor. At the same time it is possible to take necessary samples from this circulation. Circulation velocity or the reference value for flow speed (l/min) can be set from the control room.

During the process runs pH, dissolved oxygen, temperature, total volume (biomass input and process fluid outflow), as well as the gas mixing and measurement are adjusted by the process control system together with real time operating activity by the personnel on site and via 3G connection to the pilot plant. The gaseous products are recorded from the volatile outflow of the Bioreactor prior to the stabilization.

A schematic drawing of the Bioreactor is presented in Figure 1.

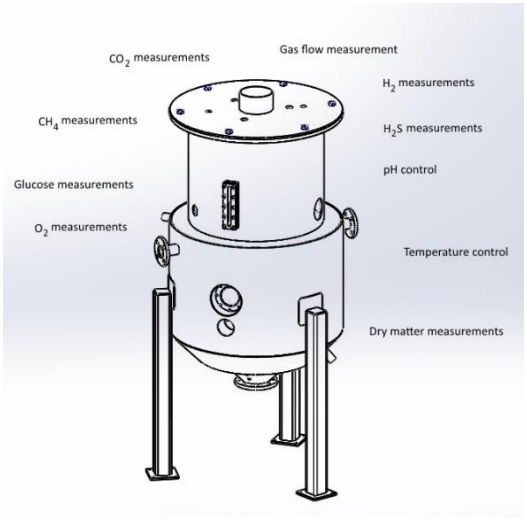


Figure 1. Schematic drawing of ABOWE Bioreactor with indication of various measurement parameters. Lecturer Anssi Suhonen, Savonia based on instructions from Adj. Prof. Elias Hakalehto, Finnoflag Oy.

4. STABILIZER was modified from a food industry boiling tank into a cooled collection unit of the bioprocess fluid containing liquid and possibly solid products of the pilot plant. In Stabilizer the temperature is decreased to 15-18 degrees Celsius from the usually much higher production temperatures in order to avoid losses in the product concentrations after the process.

The process fluid was further analyzed with Nuclear Magnetic Resonance (NMR) by Professor Reino Laatikainen at the University of Eastern Finland in Kuopio (Laatikainen *et al*, 2016). Ostfalia University of Applied Sciences in Germany provisionally experimented the downstream processing of some of the bioprocess products, under the supervision of Professor Thorsten Ahrens.

The original idea of the piloting experiments was to study the combination of gaseous, liquid and solid phases in the Bioreactor in order to produce biofuels and bioenergy, organic platform chemicals and organic fertilizers, or their raw materials. Breaking the biochemical process into bits and pieces could form this basis for any experimentation in the future. (Hakalehto *et al*, 2016b)

3 Results and Discussion

3.1 Tests in Finland

The first ABOWE testing site was Powerflute Oyj Savon Sellu fluting factory's waste water treatment plant in Kuopio, in Eastern Finland during February-March 2014. The climate conditions with temperatures between -25°C and -30°C were harsh so the functions of the pilot plant were put straight into a real "climate test". Liquids tend to get frozen in the pipes during their pumping into the pilot plant. Also the substrate for the tests, the dried sludge from the waste treatment plant

was rapidly cooling down in the piles where it was collected from the activated sludge pools. Another raw material for the experiments was the incoming waste water. In the factory there are actually two separate systems, namely the pulp factory and the corrugated carton board (fluting) machine. Waste waters from both machines arrive to the waste water treatment plant. The original raw material wood chips are treated with many chemicals during the production process. (Hakalehto *et al*, 2016b; Hakalehto *et al*, 2016c)

Products and services from the test runs at Savon Sellu were:

- Ethanol, Butanol
- 2,3-butanediol
- Organic acids
- Hydrogen
- Fertilizer biomass
- Biogas
- Purified waste water
- Decreased waste treatment expenses
- Lesser environmental and climatic load

The initial break-in test runs and international training period were carried out during the testing at Savon Sellu waste water treatment plant in January and February 2014. The first 3-4 runs were targeted for pretreating the available substrate in the pilot plant process. Both anaerobic and aerobic test runs were accomplished. In all cases, regardless of the hygienization during the hydrolysis step, the natural microflora from the activated sludge treatment process pools, especially the sulfur bacteria, contaminated the Bioreactor process broth. They were then restricted by the inoculated *Klebsiella*/*E.coli* strains which were preincubated in the Bioreactor as a nutrient bed type of inoculum. This same strategy was later used in the two anaerobic cultivations with *Clostridium sp.* (Hakalehto *et al*, 2016b) An example of the process development at the Finnish testing site is presented in Table 1.

Table 1. Description of process development outlines in the Finnish runs at the Savon Sellu fluting factory site (Hakalehto *et al*, 2016b)

	General problem	**Practical solution**	**Potential solution**
1.	Diffusion limitation	Gas flow adjustment Homogenisation with effective hydrolysis	Improved Bioreactor design for full scale plants
2.	Disturbing organisms	Speeded up inoculations Nutrient beds Hygienization of waste	Consolidated Bioprocessing
3.	Productivity problems	Mixed wastes	Simultaneous downstreaming for blocking biological down regulation
4.	Too low raw material concentration	Process fluid circulation Nutrient beds	Better pumps and valves

Microbiologically one of the most influential process chemical addition is the use of sulfuric compounds in the Savon Sellu factory process, which eventually led to the enrichment of H_2S liberating bacteria into the gas emission flow from the Bioreactor when the sludge was used as substrate. As one of the precautions for this liberation of the toxic gas flow, a specific alarm system was installed. All gases, however, were directed out of the pilot plant which completely prevented their accumulation inside, and thus the formation of occupational hazards. The general safety of the unit had been discussed beforehand with North Savo Regional Rescue Services and the authorities were satisfied with the pilot plant and the plans for its short testing period. (Hakalehto *et al*, 2016c)

The H_2S formation was on a very high level in most of the test runs. Consequently, the fed-batch function of the pilot plant was not possible to get tested during the short time-window at the Savon Sellu testing site. Furthermore, a major part of the planned time was needed for adjustments of the heating and cooling system as three of the four main process tanks were purchased as recycled industrial equipment and modified for the bioprocess use with the principle of sustainability. (Hakalehto *et al*, 2016c)

Despite the very short time schedule for the start-up phase, promising results were obtained. The primary purpose of the first tests in Finland was mainly to get familiar with the relatively complicated system of the equipment with computer control, numerous measurements, as well as temperature, pH and gas adjustments. Several sensors proved to be useful with secondary screens in the process room, besides the process control system in the control / laboratory room. (Hakalehto *et al*, 2016c)

The basic principles for steering the pilot plant were learnt, and their implementation gave promises for the potential of future bioprocess development. In the start-up phase any result from the biological multi-variable process is giving valuable information for future experiments. Improvement of the pilot plant equipment is still needed, one target being such pumps that are capable to move raw material with dry weight higher than 10-15%. Increasing this value would uplift also the yield and productivity of various biochemicals. (Hakalehto *et al*, 2016c)

On the bioprocess side, the non-aseptic principle was tested, because it could at best lower the investment costs of the process equipment to one tenth in the large production units. Therefore, the natural microflora from the waste water pools and the activated sludge caused many problems and produced overgrowth which almost took over the Bioreactor in the beginning of the experimentation. Measures for abating the background flora were attempted, but these trials were not completed during the short experimentation period. However, the nutrient bed approach with inoculation of the production

organisms prior to the major substrate addition seemed to be the correct way to tackle most of the problems caused by the background flora. (Hakalehto *et al*, 2016c)

One significant result was the production of the hydrogen gas from the waste sludge, observed earlier in the Finnoflag Oy's laboratory, which could provide 150-300 kWh of electricity daily from the Savon Sellu fluting factory's waste waters. This value was estimated from the preliminary results without any continued optimization or process development. The generation of molecular hydrogen was taking place parallel to the production of liquid chemicals, whose production levels were subjected to improvement due to the preliminary nature of the testing. The diminishment in the environmental load of Savon Sellu waste water was already comparable to the biogas process, and actually the biorefining for chemical products could be combined with cascading biogas production for the highest decrease in the climatic effects of the waste water treatment. This gas emission could be combined with methane from another reactor system in order to obtain hytane (hydrogen plus methane) gas mixture for energy production. (Hakalehto *et al*, 2016c)

3.2 Tests in Poland

The testing site in Poland was the waste management center of ZGO Gać Ltd near Wrocław, in Lower Silesia during May-early July 2014. The main substrates were potato peels from a chips factory and separately collected biowaste from households and restaurants. These substances were rather easily degradable. The potato starch had been readily degradable source of hydrolysable biomass in the experiments preceding ABOWE, carried out by Finnoflag Oy in Finland (Hakalehto *et al*, 2013). Then record levels of 2,3-butanediol productivities had been achieved (8 g/l/h). This degradative process was based on the studies with the members of Enterobacteriaceae family of facultatively anaerobic bacteria, particularly of the genus *Klebsiella* (Hakalehto, 2013; Hakalehto *et al, 2008*).

In the tests participated 24 students and six experts from Wrocław University of Technology together with Finnoflag Oy. The substrate was of a "carboxylic platform" type (den Boer *et al*, 2016a). In the experiments with sole Polish potato waste (consisting mainly of potato peels), ethanol was the principal product, besides the high amounts of hydrogen produced from the waste. Hydrogen measurement was limited to 10 000 pm due to calibration of the gas measurement unit. The hydrogen production exceeded 10 000 ppm for long periods during each run. It should be taken into account that this flow of volatiles was produced into a carrier gas flow which was not diminished in the calculations. Consequently, the levels of biohydrogen production could be considered as promising ones. (Hakalehto *et al*, 2016b)

In the beginning of the Polish testing period, heterogeneous composition of sorted biowaste was believed to disturb the process set up and control. However, this did not turn out to be a considerable problem. Instead, adding miscellaneous food waste to the process clearly boosted the production of various organic chemicals which reached a few percent of the total volume, and 15-20% of the dry weight. During these experiments the highest yields were not achieved or even tried to get achieved due to time limitations. In the future, efforts should be made to concentrating the feedstock into adequately high substrate concentrations. However, with more time and some technical upgrading of the pilot plant equipment, still much higher levels and productivities could be achieved with optimized processes. This is deducible also from the amount of unused substrate in the process residues. However, even by the current experimentation several industrial levels of biochemical productions were obtained. (Hakalehto *et al*, 2016b)

The analysis results from on-site Gas Chromatography (GC) and the Nuclear Magnetic Resonance (NMR) studies, the latter conducted by Prof. Reino Laatikainen, School of Pharmacy, University of Eastern Finland, produced somewhat different results. In some runs the levels were about 2-3 times higher in the former than the latter. This could be due to the storing of samples for the NMR in cold and transportation them to Finland, where they were analyzed much later on. It is then quite obvious that some changes could occur. Otherwise the NMR gave clear identification of the substances whereas the GC seemed to give some peaks close to each other which caused difficulties in the identification of the compounds. This was the case especially with 2,3-butanediol and valeric (pentanoic) acid. The latter was not expected to come out in the fermentation in large quantities but it was produced in high amounts. This organic acid was probably resulting from the condensation of acetic acid (two carbon molecule) and propionic acid (three carbon molecule). Both 2,3-butanediol and valeric acid could be valuable chemicals for producing butadiene (a raw material for plastics, synthetic rubber) and in cosmetic products (called also "2,3-butylene glycol"). (Hakalehto *et al*, 2016b)

Otherwise GC turned out to be a rather reliable method and it was successfully used in the pilot plant tests in Finland, Poland and Sweden. In all sampling and sample treatments it was important to separate the solids quickly enough for preventing any bacteriological activity caused degradation. The amount of products bound to the precipitating solid fraction could not be analyzed, so in future applications separation of the products in the liquid forms needs to be taken into consideration. In any case, the Polish testing period

indicated clearly the biorefinery concept's potential for producing soluble chemicals for industrial raw materials. Likewise it indicated the potential to produce hydrogen in these processes. In fact, the hydrogen production started quickly, and it was produced on remarkable levels even though this flow was integrated into the carrier gas. (Hakalehto *et al*, 2016b)

Two microbe species were applied in the Polish biorefinery process runs as additional inocula to the process broth, namely *Klebsiella mobilis* and *Escherichia coli*. These strains had been reported to metabolize glucose into organic acids, ethanol and 2,3-butanediol as a result of joint activities of their mixed cultures where a symbiotic relation between the strains developed (Hakalehto *et al*, 2008). The initial five runs were performed using potato waste as the single substrate. In these tests ethanol and acetic acids were main products. Among the gaseous products, hydrogen formation took place at elevated levels. Potato waste, however, seemed not to contain sufficient microelements to maintain generation of high level, longer chain products. It seems probable that the nitrogen and calcium levels could be limiting factors for the microbes. A significant improvement could be obtained when the kitchen biowaste were used as the initial raw material. With NMR analyses the presence of such substances as butyric acid, propionic and valeric acid at elevated levels were confirmed. Especially in the Polish Runs 7 and 8 high conversion rates to carboxylates, 0,75 and 0,81 mol/mol glucose respectively, to longer chain carboxylic acids were measured. Compared to C2–C4 carboxylic acids, longer chain carboxylic acids are superior in terms of energy content and hydrophobic nature, facilitating the downstream processing (Spirito *et al*, 2014). The results confirmed, that the biorefinery processes offer a clear advantage over conventional biowaste treatment technologies in terms of useful products which can be generated. (den Boer *et al*, 2016a)

3.3 Tests in Sweden

During the Swedish tests the pilot plant was located at a chicken farm Hagby Gård in Enköping. The farm produces roughly 800 chickens a week for slaughter (2014). The slaughterhouse, as well as their own store, is located in Västerås. (Hakalehto *et al*, 2016a; Schwede *et al*, 2017)

Finnoflag Oy led the Swedish tests which were organized by experts from Mälardalen University together with the Hagby Gård and also Savonia University of Applied Sciences. During the tests at Hagby farm, some 30% of chicken slaughterhouse waste was mixed with other biomass. The latter were chicken manure from the farm, some saw dust used as litter in the bird shelters, and occasionally some waste apples available at the farm. As additives were occasionally also used some potato flour, sugar or blueberry soup. Considerable problems emerged during the testing due to the pumps' inability to operate with the sticking feathers and the small stones originating from the bird digestion. The pumps got easily stuck with these miscellaneous particles or substances even though they had enough capacity for forwarding the biomass. Therefore, the final density and dry mass content was too low for higher product yields. However, the proof of concept was clearly demonstrated, and valuable products formed within the limits of the raw material offered. A tedious mixture of protein and lipid wastes was possible to convert first into yellowish milky broth were no particles were detected practically in overnight, and further to a solution of organic acids and alcohols. This could happen without significant loss in the dry weight of the soluble substances. (Hakalehto *et al*, 2016b)

The slaughterhouse located some 40km from the testing site so the chicken inner organs and other remaining parts were cooled for transportation. This cooling was probably not effective enough, which provided time for the mixed flora in the wastes to develop too far for the optimal substrate use in the biorefinery. In order to boost the biochemical production after the hygienization (in Hydrolyzer), strains of *Clostridium acetobutylicum* and *Clostridium butyricum* were inoculated. It is noteworthy that, even though they are generally considered as obligate anaerobes, these bacteria have been reported to withstand some oxygen occasionally. They also could stay active under 100% oxygen flow (Hell *et al*, 2010). This was used as a selective factor during the experimentation. Earlier it had been reported that the clostridial growth was boosted also by CO_2, which has been exploited for the rapid start of growth (Hakalehto, 2015; Hakalehto and Hänninen, 2012). In some runs, subsequent inoculations seemed to initiate the production of some chemicals which is implying to some quarum sensing type of signaling in the bacterial cultures. Also, addition of blueberry soup into some test runs clearly had a positive boosting effect which indicates the need for some trace elements and minerals for the best production levels. (Hakalehto *et al*, 2016b)

The expected products were short-chained organic acids, like acetate, propionate and butyrate that, if there had been more time for the test runs, could have been reduced to alcoholic and aliphatic substances. Other expected products were hydrogen and some 2,3-butanediol. Further, analysis by the NMR in Finland revealed some additional products such as valeric (pentanoic) acid and amyl alcohol. The latter was probably produced partially from the apple waste, but could get obtained also without the apples. Acetate and propionate derived from the bacteriological activity, can react with each other to form valeric acid, which also is a valuable product with a price 2-3 times of 2,3-

butanediol. (Hakalehto *et al*, 2016a; Hakalehto *et al*, 2016b)

The results show that products have been obtained both in the aerobic and anaerobic experiments. Production rates are higher when easily accessible carbohydrates and sugars are available. The highest levels of ethanol, acetic acid, propionic acid and 2,3-butanediol were obtained in the second test. Easy accessed carbohydrates in the added sugars and potato flour might explain this. Also the first run where sugar was also added showed somewhat higher product levels than later runs. (Hakalehto *et al*, 2016a)

Due to the glucose limitation of the raw material, *Klebsiella* was not effective for 2,3-butanediol production in this setting. Higher glucose levels could be facilitated by pump and mass transfer improvements, which could make it possible to gain industrial levels. Several organic acids were produced in high quantities. Improved pretreatments and elevated small carbon molecules would increase the yield also in their production.

During intensive bacteriological activity periods hydrogen production was rather high, which could give leads to development of biohydrogen production from the organic wastes, such as the animal or plant residues from the agriculture. (Hakalehto *et al*, 2016a)

Surprisingly, ABOWE tests in Sweden produced significant amounts of organic chemicals regardless of the low carbohydrate concentrations in the beginning of the experiments. This indicated the use of proteins and fats as substrates by the microbes. The microflora consisted of the natural microbiota and the added industrial strains which were managed to get function relatively well together. The yields and productivities could be increased by improvements to the process and to the equipment. (Hakalehto *et al*, 2016a; Schwede *et al*, 2017)

In an overall consideration, the Swedish testing period gave a proof of concept on a reasonable method to deal with tedious wastes from slaughterhouse and bird farm within a short time-window. This approach could serve as a significant model for the activities in many countries, where the chicken litter forms a considerable environmental problem. (Hakalehto *et al*, 2016b)

3.4 Technical Functioning of the Mobile Biorefinery Pilot Plant

During the project many challenges were faced in the designing and construction of the mobile pilot plant which would be both a real production process as well as a research laboratory. For instance, temperature and pH control for a process that is run by living organisms, brought along many questions to be solved.

The weighing of Homogenizer did not function well enough so the amount of solid matter could not be properly measured. Moreover water and other substances were added to the process from other ways than through measurements. This complicated the testing operations and the analysis thereof, as the amounts of raw materials were not always reliably known.

A lot of heat energy was needed for the hydrolysis and in the pilot plant the heating of hydrolysis functioned well. Using warm water, from a hot-water tank, as an energy source was a good solution. In an industrial scale production plant this should be recovered in cooling and used for treatment of new raw material. This means that for a viable treatment the process needs to be a continuous one and energy has to be transmitted with a heat exchanger always to new raw material. In this pilot plant this efficient energy consumption could not be implemented.

In the following figures, there are presented the measured temperature and pH values from Hydrolyzer and Bioreactor during a selected test run from all three countries. The time axis is the same in the figures of Hydrolyzer and Bioreactor.

In Figures 2 and 3 the temperature and pH measurement data from Hydrolyzer and Bioreactor during the final test run in Finland are presented.

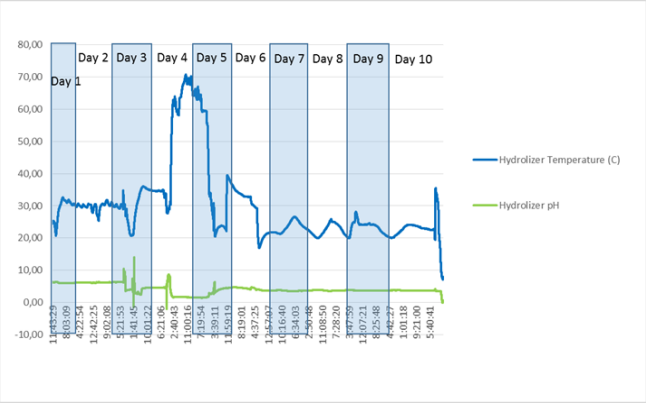


Figure 2. Hydrolyzer Temperature and pH during the Finnish Run 6 (24.3.2014 – 3.4.2014).

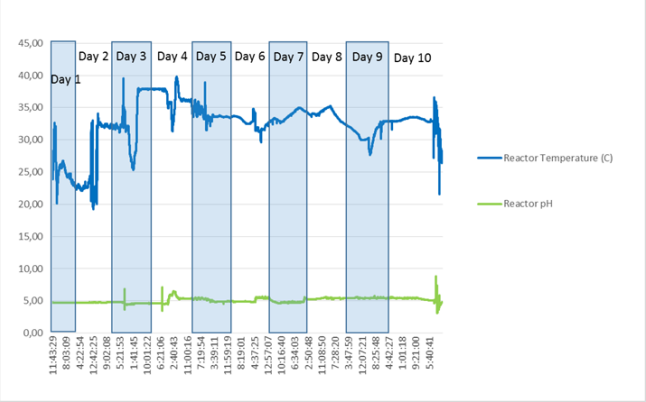


Figure 3. Bioreactor Temperature and pH during the Finnish Run 6 (24.3.2014 – 3.4.2014).

In Figures 4 and 5 the temperature and pH measurement data from Hydrolyzer and Bioreactor during Run 7 in Poland are presented.

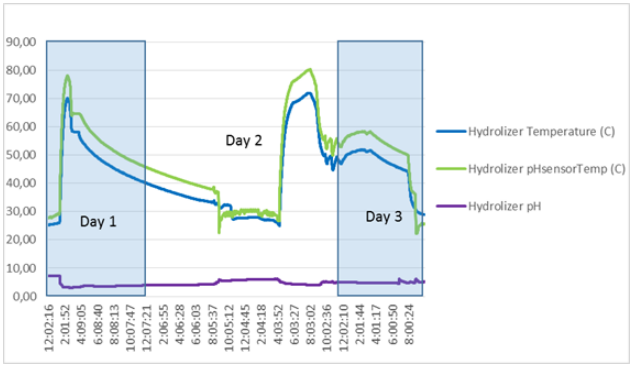


Figure 4. Hydrolyzer Temperature and pH during the Polish Run 7 (23.6.2014 – 25.6.2014).

In the Polish example, first the temperature of the Hydrolyzer had been risen up to 80°C for hygienization purposes and after that the substrate has been pumped to the Bioreactor in which the temperature had been kept as 35°C. In addition, Hydrolyzer heating had been started again, simultaneously with a bioprocess run in the Bioreactor (Day 2).

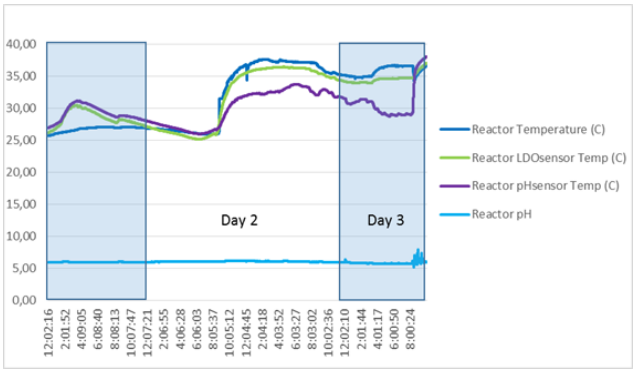


Figure 5. Bioreactor Temperature and pH during the Polish Run 7 (23.6.2014 – 25.6.2014).

In Figures 6 and 7 the temperature and pH measurement data from Hydrolyzer and Bioreactor during the final test run in Sweden are presented.

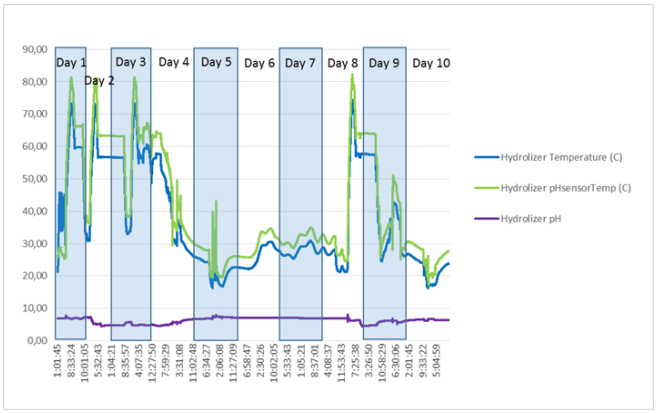


Figure 6. Hydrolyzer Temperature and pH during the Swedish Run 4 (22.9.2014 – 2.10.2014).

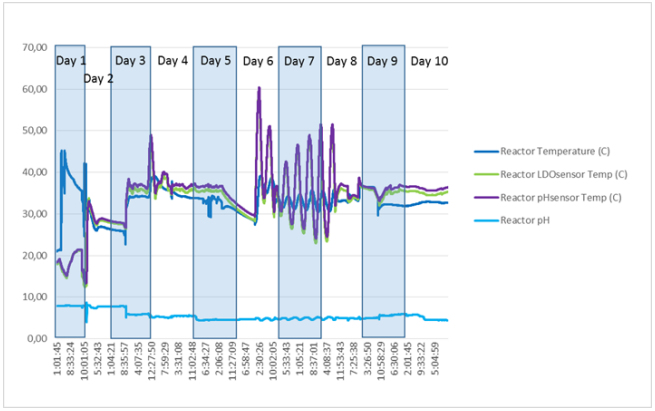


Figure 7. Bioreactor Temperature and pH during the Swedish Run 4 (22.9.2014 – 2.10.2014).

The Bioreactor was manufactured of stainless steel and for the temperature control it has a water jacket which reaches about half of the height of the Bioreactor. To the water jacket it is possible to introduce either hot water from the hot-water tank or cold water from the cooler. The objective was to gain a wide surface for the heating and cooling so that point-form high surface temperatures would not kill or hamper the bacterial strains. For the Bioreactor's temperature control these brought challenges as in principle a simple heating process changed into a process that includes two time constants. It can be concluded that the water jacket around the Bioreactor is unnecessary large, which caused too much delay for the temperature control. By speeding up the temperature measurement of exiting heating or cooling water it is possible to improve the cascading structure of the adjustment.

The Bioreactor's temperature sensor is located in the bottom where temperature cannot change very rapidly. Bioreactor mixing with airlift does not always reach the lower parts of the Bioreactor. Dissolved oxygen electrode and pH-electrode locate in the middle part of the Bioreactor and they both contain a temperature sensor. The temperature sensor of pH-electrode, however, is faster to react than the temperature sensor of

the dissolved oxygen electrode. pH-electrode has a considerably smaller mass and in addition, the tip of the electrode, in which the temperature sensor is locating, is narrow. Hence the values that these two temperature sensors are showing might differ from each other.

PH adjustment of the Bioreactor was known to be challenging and even more efforts should have been put to that control. However, pH has remained surprisingly stable. Temperature compensation of the pH electrode rectifies the change of the current from pH electrode that is due to temperature, but also the pH of a substance is changing as temperature changes and that change is not linear.

The emptying of the Bioreactor did not function well. Continuous recovery of products was not tried because the Stabilizer was not insulated and the risk of emitted H_2S from the process broth could not be handled.

4 Conclusions

A lot of challenges were faced in ABOWE because of the very tight time schedule for engineering and construction of the mobile biorefinery pilot plant. Besides, there were only two months available for test runs in each of the three countries. There came out some problems which probably were due to insufficient time to implement all planned settings before the test runs had to be started. Planned systems were not always put completely into operation because of their deficiencies or because there were not time for training. This led to situation where the major parts of the system were operated manually although automatic functions had been designed.

From the microbiology side, the combinations of the natural microbial flora with some added microbial strains should be further tested in future. It would be necessary to continue the experiments with the cellulosic waste in order to demonstrate the full potential of the biorefinery principle. It could then also be possible to combine the control with the activities of the natural flora into an overall bioprocess steering.

In ABOWE Polish tests pure cultures of microbes were used. However, volatile fatty acids (VFA) can be produced from various organic wastes also with mixed cultures under anaerobic conditions (Agler *et al*, 2011). According to Agler *et al.* (2011) and Spirito *et al.* (2014), undefined mixed cultures (or unrestricted microbiomes) offer a clear advantage over pure cultures because open microbiomes can, due to the metabolic flexibility, tolerate the complexity and variability of substrates. Moreover, sterilization and aeration can be eliminated in many cases since undefined mixed cultures can grow under non-aseptic conditions and often in anaerobiosis, too. (den Boer *et al*, 2016b)

More test runs would be needed for fully optimizing the effects of various parameters to further improve the yield of the biorefining. Gaining and maintaining the optimal conditions for bioprocesses is demanding as many raw materials and their concentrations or the products can affect even radically to the process and its velocity. From the adjustment technology point of view it is the question of a multi-parameter process and in addition the events are nonlinear.

It is possible to compile a mathematical model of the Bioreactor that can be simulated. The parameters have to be known as well as the interrelations between parameters. Based on measurement data, the Bioreactor operations could be modelled and transfer functions could be gained. Then various adjustment solutions should be simulated with a numerical computing program. For example, it would be possible, with logging data from automation system with regular intervals, to further model the temperature behavior of the Bioreactor such as temperature measurement and adjustment as well as to try cascade adjustment. To calculate heating and cooling power there should be, in addition to temperature measurement, also flow measurement in place. In measurements-based modelling, the adjustment should be done manually so that the effect of steering would be seen in the responses from the measurements reliably enough.

Implementation of degradative and recycling function of the Nature's microbiota into industrial applications is the leading principle in the Finnoflag Oy's biorefinery technology. This requires understanding of the interactions between the biomass (whose composition is subjected to variations), its natural flora, and the added strains and enzymes. The main purpose of the biorefinery pilot plant experiments was to give a reliable proof of concept on the industrially important substances producible in a sustainable way. This goal was achieved successfully, and several overall challenges were overcome during the testing in three countries. (Hakalehto *et al*, 2016b)

According to the analysis on the climatic impacts of the biotechnological processes, a combined production strategy including both 1. biorefining of chemicals from biomasses and 2. biogas process based on its residues could add value, if technologies were applied together with a cascading principle. This approach could also bring along an effective solution for eliminating the waste problems. In this case the biorefining and the preliminary hydrolysis should take place preferably in a consolidated bioprocess (CBP) where the waste macromolecules would be hydrolyzed simultaneously with the actual upstream process (Hakalehto, 2015). In case of the pilot plant the hydrolysis was partially going on also after the transfer of the pretreated substrate from the Hydrolyzer into the Bioreactor. Fast moving of the broth, from where the biochemicals were collected using the CBP principle, with a subsequent transfer of the biomass from the biorefinery into a biogas production unit, could optimally contribute in the

lowering of any climatic effect of the waste treatment. Then the biogas process could be boosted by the remaining organic acids in the solid fraction. Also the uplifted biochemical and gas production levels after optimization of the piloting and scale up trials would produce improvements in greenhouse gas reduction. (Hakalehto, 2015; Hakalehto *et al*, 2016b)

den Boer *et* al. (2016b) define some potential configurations of the biorefinery process in the waste management system and synergies of co-location with other waste handling plants. Scenario 1 is a biorefinery combined with a municipal incineration plant for residual waste. Separate collection and treatment of biowaste improves the calorific values of residual waste. Building biorefinery, as a central biowaste recycling plant next to an incineration plant would enable heat generated in the incineration to be utilized for heating up the biorefinery process. Scenario 2 co-locates a biorefinery with a biogas plant within one regional biowaste treatment plant. The biorefinery would add value in the pretreatment of separately collected biowaste by first processing high value commercial products from it. The residual biodegradable fraction of biorefinery process could be further fed to a biogas plant in order to produce energy. The benefit from combining the biorefinery and biogas plant is that the excess heat of the biogas plant could be used to heat up the biorefinery process. Scenario 3 considers locating biorefinery in combination with an industrial wastewater treatment plant. Biorefinery technology could be implemented at a range of food processing industries such as a potato chips and snacks factory. Nowadays, the large quantities of biowaste that the industry generates need to be transported to a remote biogas plant. The benefit would be to include the biorefinery process directly on-site, which saves in the transportation costs. The plant would also benefit from diversification of products they generate and options for new business. For the treatment of the residues of the biorefinery process, the optimal solution would be to utilize existing wastewater treatment plant. Food industry plants normally generate large amounts of wastewater which needs to be treated directly on-site. The most common method to treat wastewater sludges is wet digestion. Here, the co-digestion of wastewater sludges and residues from the biorefinery process would be the most beneficial solution. The heat generated in the digestion plant could be again utilized for the biorefinery process. Stabilized sludges from the wet digestion could be used as high quality fertilizer to close the natural cycle of nutrients. (den Boer *et al*, 2016b)

Acknowledgements

The ABOWE project was funded by the European Union (European Regional Development Fund). Construction of the mobile biorefinery pilot plant was co-funded by the Ministry of Employment and Economy in Finland and the Regional Council of Pohjois-Savo, Finland.

References

M.T. Agler, B.A. Wrenn, S.H. Zinder, and L.T. Angenent. Waste to bioproduct conversion with undefined mixed cultures: the carboxylate platform. Special Issue – Applied Microbiology, *Trends in Biotechnology*, 29(2), 2011.

G. Bastin and D. Dochain, On-line estimation and adaptive control of bioreactors. *Process Measurement and Control 1*. Amsterdam: Elsevier Science Publishers B.V., 1990.

E. den Boer, A. Łukaszewska, W. Kluczkiewicz, D. Lewandowska, K. King, A. Jääskeläinen, A. Heitto, R. Laatikainen, and E. Hakalehto. Biowaste conversion into carboxylate platform chemicals. In: E. Hakalehto (ed.) *Microbiological Industrial Hygiene*. New York, USA: Nova Science Publishers, Inc., 2016a.

E. den Boer, A. Łukaszewska, W. Kluczkiewiczc, D. Lewandowska, K. King, T. Reijonen, T. Kuhmonen, A. Suhonen, A. Jääskeläinen, A. Heitto, R. Laatikainen, and E. Hakalehto. Volatile fatty acids as an added value from biowaste. *Waste Manag*., 58: 62-69, 2016b.

E. Hakalehto (ed.) *Alimentary microbiome - a PMEU Approach*. New York, USA: Nova Science Publishers, Inc., 2012.

E. Hakalehto. Interactions of Klebsiella sp. with other intestinal flora. In: L.A. Pereira and A. Santos (eds.) *Klebsiella infections: Epidemiology, pathogenesis and clinical outcomes*. New York, USA: Nova Science Publishers, Inc., 2013.

E. Hakalehto. Enhanced microbial process in the sustainable fuel production. In: J. Yan (ed.) *Handbook of clean energy systems*. Chichester, UK: Wiley JR & Sons. Inc, 2015.

E. Hakalehto and O. Hänninen. Gaseous CO2 signal initiate growth of butyric acid producing Clostridium butyricum both in pure culture and in mixed cultures with Lactobacillus brevis. *Can J Microbiol*, 58(7): 928-931, 2012.

E. Hakalehto, T. Humppi, and H. Paakkanen. Dualistic acidic and neutral glucose fermentation balance in small intestine: Simulation in vitro. *Pathophysiology*, 15(4): 211-220, 2008.

E. Hakalehto and L. Heitto. Minute microbial levels detection in water samples by Portable Microbe Enrichment Unit technology. *Environment and Natural Resources Research*, 2(4): 80-88, 2012.

E. Hakalehto, A. Jääskeläinen, T. Humppi, and L. Heitto. Production of energy and chemicals from biomasses by micro-organisms. In: E. Dahlquist (ed.) *Biomass as energy source: resources, systems and applications*. London, UK: CRC Press, Taylor & Francis Group, 2013.

E. Hakalehto, A. Heitto, H. Andersson, J. Lindmark, J. Jansson, T. Reijonen, A. Suhonen, A. Jääskeläinen, R. Laatikainen, S. Schwede, P. Klintenberg, and E. Thorin. Some remarks on processing of slaughterhouse wastes from ecological chicken abattoir and farm. In: E. Hakalehto (ed.) *Microbiological Industrial Hygiene*. New York, USA: Nova Science Publishers, Inc., 2016a.

E. Hakalehto, A. Heitto, A. Suhonen, and A. Jääskeläinen. ABOWE project concept and Proof of Technology. In: E.

Hakalehto (ed.) *Microbiological Industrial Hygiene*. New York, USA: Nova Science Publishers, Inc., 2016b.

E. Hakalehto, A. Heitto, H. Niska, A. Suhonen, R. Laatikainen, L. Heitto, E. Antikainen, and A. Jääskeläinen. Forest industry hygiene control with reference to waste refinement. In: E. Hakalehto (ed.) *Microbiological Industrial Hygiene*. New York, USA: Nova Science Publishers, Inc., 2016c.

M. Hell, C. Bernhofer, S. Huhulescu, A. Indra, F. Allerberger, M. Maass, and E. Hakalehto. How safe is colonoscope-reprocessing regarding Clostridium difficile spores? *The Journal of Hospital Infection*, Vol. 76, Supplement 1: Abstracts, 8th International Congress of the Hospital Infection Society, 10-13 October 2010, Liverpool, UK, S21-22

A. Jääskeläinen and E. Hakalehto, *ABOWE and Beyond - Baltic Sea Biorefinery Piloting 2014. ABOWE Biorefinery Final Summary Report*. Kuopio, Finland: Savonia University of Applied Sciences, 2015. Available at: http://portal.savonia.fi/amk/sites/default/files/pdf/eng/abowe/ABOWE_Biorefinery_Final_Summary_Report

A. Jääskeläinen and E. Hakalehto. Biorefinery education as a tool for teaching sustainable development. In: W. Leal Filho (ed.) *Implementing Sustainability in the Curriculum of Universities*. *Approaches, Methods and Projects.* Cham, Germany: Springer International Publishing, 2018.

R. Laatikainen, P. Laatikainen, and E. Hakalehto. Quantitative quantum mechanical nmr analysis: The superior tool for analysis of biofluids. In: Proceedings of the 1st Int. Electron. Conf. Metabolomics, 1–30 November 2016. *Sciforum Electronic Conference Series*, 1, C005. doi:10.3390/iecm-1-C005

C.M. Spirito, H. Richter, K. Rabaey, A.J.M. Stams, and L.T. Angenent. Chain elongation in anaerobic reactor microbiomes to recover resources from waste. *Current Opinion in Biotechnology*, 27: 115-122, 2014.

S. Schwede, E. Thorin, J. Lindmark, P. Klintenberg, A. Jääskeläinen, A. Suhonen, R. Laatikainen, and E. Hakalehto. Using slaughterhouse waste in a biochemical based biorefinery -results from pilot scale tests. *Environmental Technology*, 10: 1275-1284, 2017.

Modelling of Target-Controlled Infusion of Propofol for Depth-of-Anaesthesia Simulation in Matlab-Simulink

Gorazd Karer

Faculty of Electrical Engineering, University of Ljubljana, Slovenia, gorazd.karer@fe.uni-lj.si

Abstract

Total intravenous anaesthesia (TIVA) is an anaesthesiologic technique, where substances are injected intravenously. The anaesthesiologist adjusts the injection of intravenous anaesthetic agents regarding the depth of anaesthesia. In the paper, we present a model of an anaesthetic agent, namely propofol, influencing the depth of anaesthesia. The influence of propofol is linked to the concentration of the drug in the appropriate compartment. First, the modelling of pharmacokinetics of propofol is introduced. The 3-compartmental model and the effect-site model are presented, the relevant model parameters are given. Next, the model is verified by comparing the simulation results to the data file that was recorded by the Orchestra Base Primea infusion workstation during a medical procedure, which lasted about 40 minutes. The simulation results are presented and the predictive quality of the model is evaluated.

The presented model for Matlab-Simulink provides a basic tool for further researching the dynamics of anaesthetic depth. Despite the fact that more data must be obtained in order to properly validate the model, the presented model provides a basis for running simulations and testing various scenarios of propofol administration and is usable for developing and testing closed-loop control approaches for automatic control of depth of anaesthesia.

Keywords: target-controlled infusion, Propofol, depth-of-anaesthesia, Matlab-Simulink

1 Introduction

To perform a general anaesthesia, it is necessary to use substances, which enable deep unconsciousness, analgesia, amnesia and muscle relaxation, all required for performing a surgery or a diagnostic procedure. General anaesthesia and related dynamic activities in the human body is a complicated process, which includes pharmacokinetic and pharmacodynamic mechanisms, which have not been fully studied yet.

During the general anaesthesia, the anaesthesiologist needs to monitor the patient's vital functions and maintain the functions of vital organs. To achieve anaesthesia, substances are introduced in different manners into the patient's body. In clinical practice, the most commonly used methods are the intravenous induction of an anaesthetic agent, i.e., injection of the anaesthetic into a vein, and inhalation induction of anaesthesia, whereby the patient inhales the substance from the breathing mixture. Total intravenous anaesthesia (TIVA) is an anaesthesiologic technique, where substances are injected intravenously.

The anaesthesiologist needs to adjust the dosage of anaesthetic to maintain the appropriate depth of general anaesthesia according to pharmacokinetics and pharmacodynamics of the anaesthetic agent and considering the type of procedure. Inadequate depth of anaesthesia is manifested with the activation of sympathetic nerves or in the most unlikely event with the patient awakening. Too deep anaesthesia is manifested with a drop in blood pressure level and heart rate frequency as well as slow postoperative awakening of the patient from general anaesthesia. In modern clinical practice, the depth of anaesthesia is determined by assessing the relevant clinical signs (iris, sweating, movements), by interpreting hemodynamic measurements (Potočnik et al., 2011) and by estimating the depth of anaesthesia from EEG signals, for which several established measurement systems already exist, e.g. BIS index, Narcotrend, Scale Entropy and Response Entropy. BIS index measurement is a non-invasive method, where a BIS monitor is connected to electrodes on the patient's head. By measuring the EEG signals the bispectral index is defined, representing the depth of anaesthesia. The BIS monitor provides a single dimensionless number, which ranges from 0 (equivalent to EEG silence) to 100. A BIS value between 40 and 60 indicates an appropriate level for general anaesthesia, whereas for long-term sedation due to head injuries a value below 40 is appropriate. The reference can thus be set to the applicable value; the manner and speed of approaching the reference value depend on the specific characteristics of the procedure and the pharmacokinetics and pharmacodynamics of the substance in the patient's body.

The problem of modelling the effect of propofol is described in literature in various ways. For such purposes, pharmacokinetic and pharmacodynamic models have been developed, such as in (Marsh et al., 1991; Schnider et al., 1998, 1999; Kataria et al., 1994; Schüttler and Ihmsen, 2000; Kenny and White, 1990). The models typically define the basic structure of the dynamic operating system of propofol and the parameters depend on individual patients. The values of model's parameters are affected by the patient and his characteristics (weight, height, age, sex etc.) as well as individual sensitivity to propofol and the

ability to excrete propofol.

Several developed pharmacokinetic models are used in certain infusion pumps for target controlled infusion (TCI), where the pump sets the proper flow of the medication with regard to the model. The problem with these models is that they often do not reflect the real dynamics, which also depends on individual sensitivity of the patients to the substance, therefore such approaches, based on open-loop induction, often do not yield the best performance.

In the paper, we present a dynamical model of propofol influencing the depth of anaesthesia, which is connected to the concentration of the drug in the appropriate compartments. The paper is organized as follows. First, modelling of pharmacokinetics of propofol is introduced. The 3-compartmental model and the effect-site model are presented, the relevant model parameters are given. Next, the model is verified by comparing the simulation results to the data file of an actual anaesthetic application of target-controlled infusion of intravenously administered propofol, which was recorded by the Orchestra Base Primea infusion workstation during a medical procedure that lasted about 40 minutes. The simulation results are presented and the predictive quality of the model is evaluated. Finally, we give some concluding remarks.

2 Modelling the pharmacokinetics of propofol

2.1 The 3-compartmental model

The pharmacokinetics of the derived model is based on the Marsh model (Marsh et al., 1991). The dynamic relations are based on a well-established 3-compartmental model structure, as shown in Figure 1.

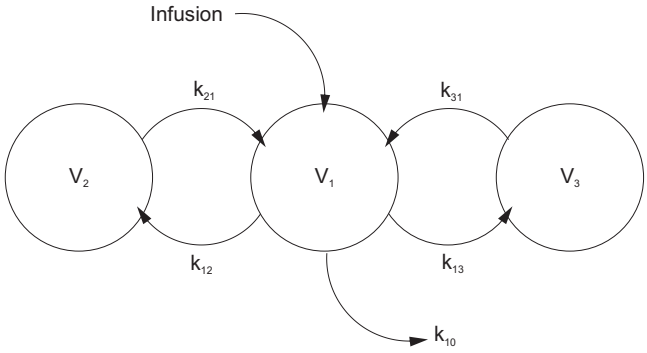


Figure 1. The pharmacokinetics of the 3-compartmental model.

The 3-compartmental model can be described as follows:

- The drug (namely *propofol*) is injected intravenously into the central compartment (V_1), representing the blood (or plasma) in the body - contained primarily in the arteries and veins and the directly influenced tissues and organs, such as brain, heart, liver, kidney etc.

- The second compartment (V_2) represents the group of tissues that are indirectly affected by the amount of drug in the central compartment, i.e., mainly the muscles. The exchange of the drug with the central compartment is denoted by k_{12} and k_{21}.

- The third compartment (V_3) represents the group of tissues that can store a certain amount of drug, but the exchange with the central compartment is rather slow, i.e., mainly the fat. However, the amount of drug in these tissues influences the amount of the drug in the central compartment in the long run. The exchange of the drug with the central compartment is denoted by k_{13} and k_{31}.

- The drug is eliminated from the body with a rate denoted by k_{10}.

The internal dynamics of the model can be formulated by using

$$\frac{dx_1}{dt} = \phi - k_{12}x_1 - k_{13}x_1 - k_{10}x_1 + k_{21}x_2 + k_{31}x_3 \quad (1)$$

$$\frac{dx_2}{dt} = -k_{21}x_2 + k_{12}x_1 \quad (2)$$

$$\frac{dx_3}{dt} = -k_{31}x_3 + k_{13}x_1 \quad (3)$$

where the variables x_1, x_2, and x_3 represent the amount of the drug in compartment V_1, V_2, and V_3. respectively. The infusion flow rate is denoted as ϕ. As noted above, the parameters k_{12}, k_{21}, k_{13}, and k_{31}, represent the partition coefficients that determine the speed at which the drug goes from one particular compartment to another. Finally, k_{10} is the rate of elimination of the drug from the body.

Note that the concentration in the central compartment is often referred to as plasmatic concentration.

2.2 Effect-site concentration model

The effect site for the drug propofol is basically the central nervous system. The effect site is thus part of the central compartment, but the effect of the drug is subject to some dynamics with regard to the (theoretical) concentration in the central compartment. This is mainly due to transportation delay as the drug concentration in the central compartment is not homogenous, which is evident especially during the transient response, i.e., when the amount of the drug in the central compartment is changing rapidly. The effect-site concentration is therefore a representation of a volumeless 4th compartment, where the drug is active. This compartment is virtually linked to the central compartment.

Therefore, a 1st order model has been used to describe the effect-site concentration dynamics, as given in

$$\frac{dx_e}{dt} = -k_{e0}x_e + k_{e0}x_1 \quad (4)$$

where the virtual link between the central and the effect-site compartment is characterised by the coefficient k_{e0}, which is actually the inverse time constant of the dynamic system describing the connection between the plasmatic concentration and the effect-site concentration x_{e0}.

The schematics explaining the effect-site concentration dynamics are presented in in Figure 2.

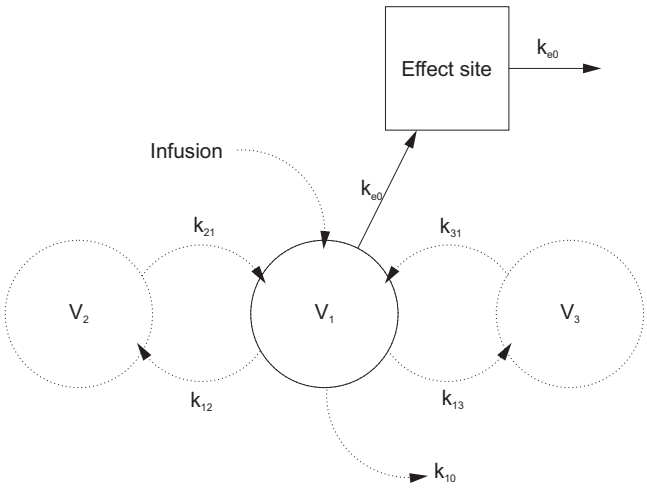


Figure 2. The effect-site concentration dynamics.

2.3 Model parameters

The parameters of the model are taken from (Marsh et al., 1991) and (Orc). The values are given in Table 1.

Table 1. Parameter values.

Parameter	Value
V_c	0.228 l/kg
k_{10}	0.119/min
k_{12}	0.112/min
k_{13}	0.0419/min
k_{21}	0.055/min
k_{31}	0.0033/min
k_{e0}	1.21/min

When developing the model in Matlab-Simulink we have to consider the units and the dilution factors of the drug. For example, the parameters expressed in min^{-1} have to be converted to s^{-1}. Furthermore, the input flow of propofol is typically expressed in ml/h, and the dilution is $10mg/ml$, i.e., 1%. The output concentration (plasmatic and effect-site) are expressed in $\mu g/ml$.

3 Model verification

The Marsh model is also used in the infusion workstation *Orchestra Base Primea* (produced by *Fresenius Kabi*) (Orc). We obtained a textual output data file of an actual anaesthetic application of target-controlled infusion of intravenously administered propofol, which was recorded by the Orchestra Base Primea infusion workstation during a medical procedure that lasted about 40 minutes. The data recorded in the output file are as follows:

- The infusion flow of propofol.
- The predicted plasmatic concentration of propofol c_p.
- The predicted effect-site concentration of propofol c_e.
- Important events (alarms, occlusions, syringe changes, target-value changes).
- Patient data:
  - age (43 years),
  - weight (78 kg),
  - height (177 cm), and
  - gender (male).

A Matlab-based parser was developed, which is able to processes the textual output data files recorded by the Orchestra Base Primea infusion workstation so as to obtain suitably formatted time-stamped data arrays. In such a manner, the obtained data arrays can easily be used in the Matlab-Simulink environment.

4 Simulation results

4.1 Propofol inflow

We used the presented model for simulating the system behaviour with regard to the response to the inflow of propofol influencing the depth of anaesthesia through plasmatic concentration and effect-site concentration of propofol. The simulated inflow of propofol was adjusted according to the data recorded by the Orchestra Base Primea infusion workstation.

The inflow-signal of propofol $\phi_{propofol}$ was parsed from the recorded data and is shown in Figure 3.

Note that first a bolus-dose is administered so as to rapidly increase the concentration of propofol in the body. This phase is called the induction of anaesthesia and results in the patient losing consciousness. Later, a suitable dose of propofol is continuously administered in order to keep the proper anaesthetic depth. A close-up of the second transient phase is shown in Figure 4.

The presented Matlab-Simulink model was fed the propofol-inflow signal $\phi_{propofol}$ and the resulting trajectories of propofol concentration in the central compartment and in the effect-site compartment (c_p and c_e, respectively) were compared to the data recorded by the Orchestra Base Primea infusion workstation.

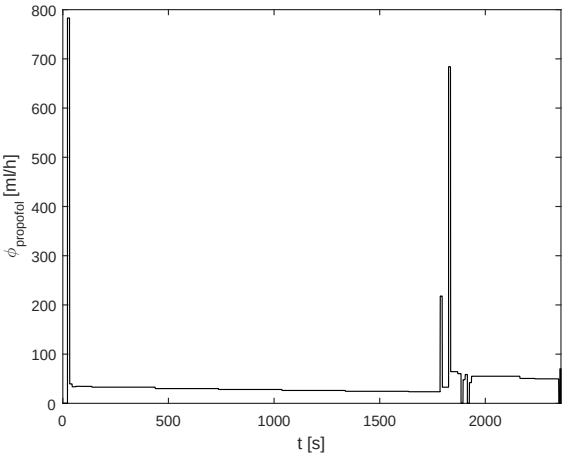


Figure 3. The inflow of propofol $\phi_{propofol}$.

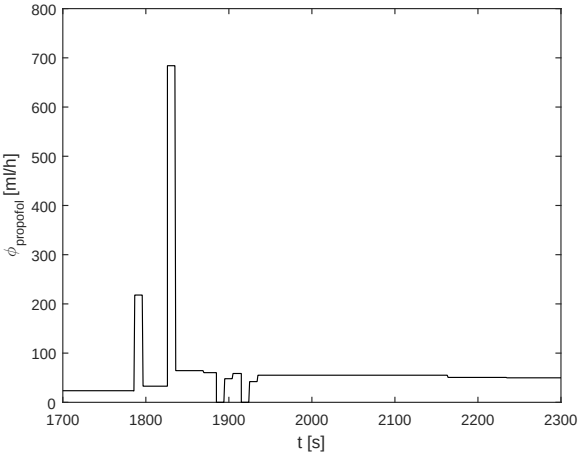


Figure 4. The inflow of propofol $\phi_{propofol}$ (close-up).

4.2 Plasmatic concentration

The simulated plasmatic concentration of propofol $c_{p,sim}$ trajectory is shown in Figure 5. The simulated results are compared to the parsed data from the Orchestra Base Primea infusion workstation data file $c_{p,data}$. A close-up of the second transient phase is shown in Figure 6.

4.3 Effect-site concentration

Similarly, the simulated effect-site concentration of propofol $c_{e,sim}$ trajectory is shown in Figure 7. The simulated results are compared to the parsed data from the Orchestra Base Primea infusion workstation data file $c_{e,data}$. A close-up of the second transient phase is shown in Figure 8.

4.4 Evaluation of the predictive quality of the model

In order to evaluate the predictive quality of the model, the simulated results are compared to the parsed data from the Orchestra Base Primea infusion workstation data file. Some established quantitative measures for predictive

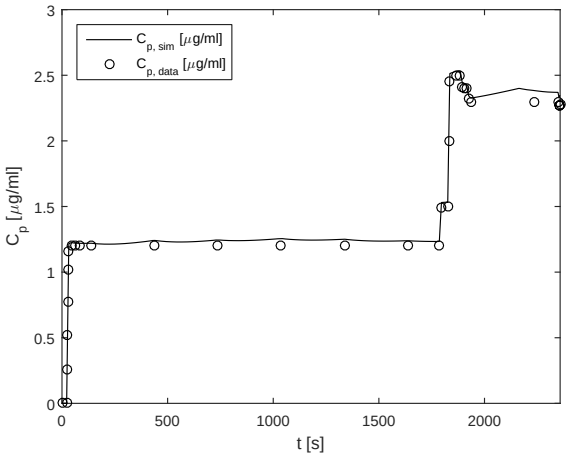


Figure 5. The simulated plasmatic concentration of propofol $c_{p,sim}$ compared to the parsed data $c_{p,data}$.

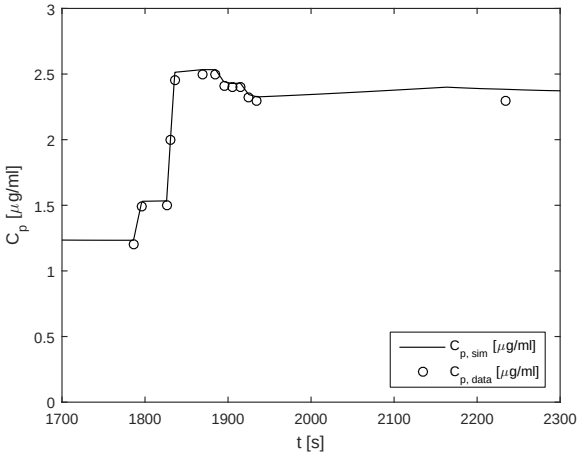


Figure 6. The simulated plasmatic concentration of propofol $c_{p,sim}$ compared to the parsed data $c_{p,data}$ (close-up).

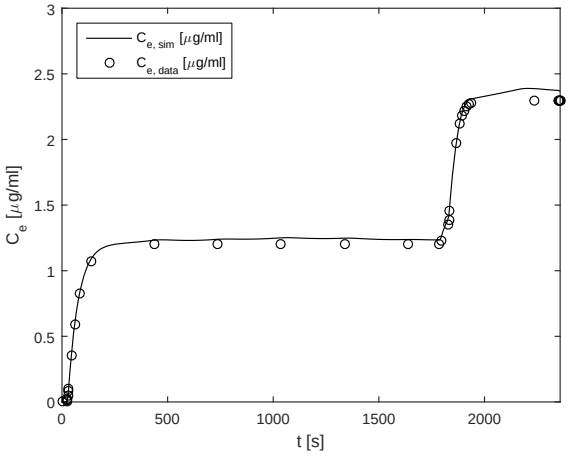


Figure 7. The simulated effect-site concentration of propofol $c_{e,sim}$ compared to the parsed data $c_{e,data}$.

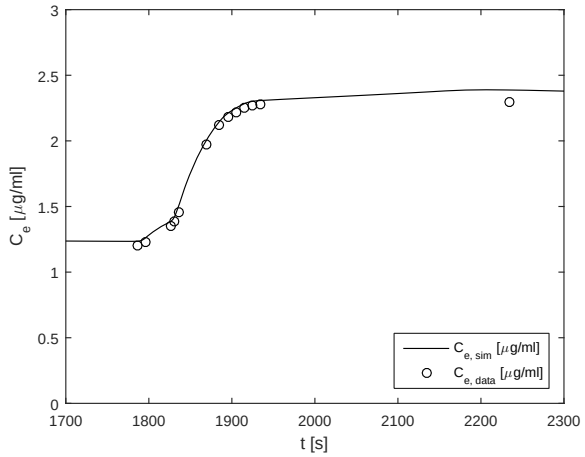


Figure 8. The simulated effect-site concentration of propofol $c_{e,sim}$ compared to the parsed data $c_{e,data}$ (close-up).

quality are *prediction mean square error* (PMSE), *median performance error* (MDPE), and *median absolute performance error* (MDAPE) (Mertens et al., 2003). The aforementioned measures are calculated as defined in

$$PMSE_x = \frac{1}{N}\sum_{i=1}^{N}(x_{i,sim} - x_{i,data})^2 \quad (5)$$

$$MDPE_x = median\{\frac{x_{i,data} - x_{i,sim}}{x_{i,sim}} \cdot 100\%\}_{i=1,\ldots,N} \quad (6)$$

$$MDAPE_x = median\{|\frac{x_{i,data} - x_{i,sim}}{x_{i,sim}}| \cdot 100\%\}_{i=1,\ldots,N} \quad (7)$$

where x_{sim} and x_{data} stand for the simulated and the "measured" data, respectively, and N is the number of data points in the data set.

In our case, the values of the aforementioned measures are presented in Table 2.

Table 2. Predictive quality measures.

Signal x	$PMSE_x$	$MDPE_x$	$MDAPE_x$
C_p	$2.21 \cdot 10^{-3}$	-1.67%	2.00%
C_e	$1.62 \cdot 10^{-3}$	-2.35%	2.78%

As the Orchestra Base Primea infusion workstation suffers from some error when logging the propofol-flow data $\phi_{propofol}$, it is sensible to take into account the final cumulative amount of the administered drug. In this case, the total amount of propofol used was 25.5 ml, whereas the simulated consumption was 26.2 ml. It is clear that the simulated propofol concentration signals, both plasmatic and effect-site, are influenced considerably by the aforementioned error.

In general, the propofol-flow $\phi_{propofol}$ error is generally not uniformly distributed. Nevertheless, the simulation results can be improved by simply pondering the simulated propofol inflow by a factor of $\frac{25.5ml}{26.2ml}$. The newly simulated results are again compared to the parsed data from the Orchestra Base Primea infusion workstation data file in Figures 9 and 10 for plasmatic and effect-site concentration trajectories, respectively.

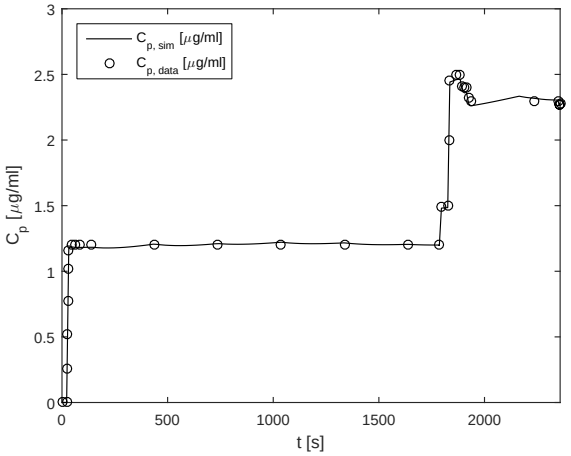


Figure 9. The newly simulated plasmatic concentration of propofol $c_{p,sim}$ compared to the parsed data $c_{p,data}$.

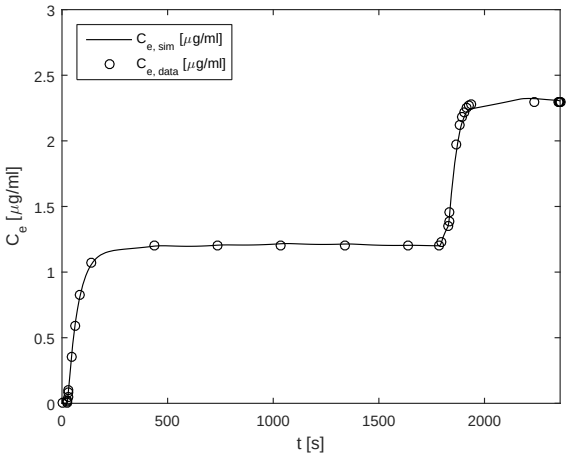


Figure 10. The newly simulated effect-site concentration of propofol $c_{e,sim}$ compared to the parsed data $c_{e,data}$.

In this manner, the predictive quality measures are improved by a considerable amount, as shown in Table 3.

Table 3. Predictive quality measures (modified).

Signal x	$PMSE_x$	$MDPE_x$	$MDAPE_x$
C_p	$0.65 \cdot 10^{-3}$	1.16%	1.38%
C_e	$0.32 \cdot 10^{-3}$	0.324%	1.14%

5 Conclusions

The developed model for Matlab-Simulink provides a basic tool for further researching the dynamics of depth of

anaesthesia. The model enables improvement and refinement of model quality for propofol for assessment of the dynamic properties of anaesthetic depth. However, more data must be obtained in order to properly validate the model. We will work with the anaesthesiologic team of the University clinical centre in Ljubljana in order to reach this goal. In addition, the measured data from real medical procedures will enable further refinement of the model parameters and possibly a structural improvement of the considered dynamical relations.

That said, the presented model provides a basis for running simulations and testing various scenarios of propofol administration that could lead to better administration protocols due to a deeper insight of the mechanisms of depth of anaesthesia.

Finally, the model is usable for developing and testing closed-loop control approaches for automatic control of anaesthetic depth, which will be the main direction of our joint future research in collaboration with the anaesthesiologic team of the University clinical centre in Ljubljana.

References

Operator's Guide: Infusion Workstation: Orchestra Base Primea.

B. K. Kataria, S. A. Ved, H. F. Nicodemus, G. R. Hoy, D. Lea, M. Y. Dubois, J. W. Mandema, and S. L. Shafer. The pharmacokinetics of propofol in children using three different data analysis approaches. *Anesthesiology*, 80:104–122, 1994.

G. N. Kenny and M. White. Intravenous propofol anaesthesia using a computerised infusion system. *Anaesthesia*, 46:204–209, 1990.

B. Marsh, M. White, N. Morton, and G. N. Kenny. Pharmacokinetic model driven infusion of propofol in children. *Br J Anaesth*, 67:41–48, 1991.

M. J. Mertens, F. H. M. Engbers, A. G. L. Burm, and J. Vuyk. Predictive performance of computer-controlled infusion of remifentanil during propofol/remifentanil anaesthesia. *Br J Anaesth*, 90(2):132–141, 2003.

I. Potočnik, V. Novak Janković, T Štupnik, and B. Kremžar. Haemodynamic changes after induction of anaesthesia with sevoflurane vs. propofol. *Signa Vitae*, 6(2):52–57, 2011.

T. W. Schnider, C. F. Minto, P. L. Gambus, C. Andresen, D. B. Goodale, S. L. Shafer, and E. J. Youngs. The influence of method of administration and covariates on the pharmacokinetics of propofol in adult volunteers. *Anesthesiology*, 88: 1170–1182, 1998.

T. W. Schnider, C. F. Minto, S. L. Shafer, P. L. Gambus, C. Andresen, D. B. Goodale, and E. J. Youngs. The influence of age on propofol pharmacodynamics. *Anesthesiology*, 90:1502–1516, 1999.

J. Schüttler and H. Ihmsen. Population pharmacokinetics of propofol: a multicenter study. *Anesthesiology*, 92(3):727–738, 2000.

Development of a Genetic Algorithms Optimization Algorithm for a Nutritional Guidance Application

Petri Heinonen[1] Esko K. Juuso[2]

[1]Nutri-Flow Oy, Finland, petri.heinonen@nutri-flow.fi
[2]Control Engineering, Faculty of Technology, University of Oulu, Finland, esko.juuso@oulu.fi

Abstract

Personalized easy to follow nutritional guidance is getting more important, since lifestyle related health problems are increasing. To gain a healthy balanced diet usually requires knowledge of a licensed nutritionist. There is a Fuzzy Expert System (FES) which applies knowledge of nutritionists, health data of an individual, personalized nutritional recommendation, and a meal diary with food composition data to balance a diet. FES generates a set of foods and beverages which should be altered in the diet with information on the direction and importance of the change. This paper presents a selection and a development of an optimization algorithm to be integrated with FES to provide easy to follow nutritional guidance. The selection process is carried out as a literature review. The development of selected Genetic Algorithms (GA) is carried out as an integrated part of Nutritional Guidance application, Nutri-Flow®, since FES generates the search space, and is an important part of a Fitness Function of the optimization algorithm. The selection of the design parameters, are described and the test results are presented. Validation of the overall model is carried out with an expert analysis and comparison of the nutrient intake from the initial diet and recommended diet.

Keywords: genetic algorithms, optimization, nutritional guidance

1 Introduction

Personal dietary guidance is an important tool for achieving global and national targets of the battle on non-communicable diseases caused by lifestyle habits on diet and physical exercise (Heinonen, 2009). Micronutrient malnutrition due to eating habits is getting common while non-nutrient dense energy rich foods are getting common in diets (IFPRI, 2016).

The Internet is filled with calorie calculators and other similar applications which calculates energy and nutrient intake levels according to a meal diary. Average consumers cannot balance their diets by knowing which micronutrients should be added to balance the diet. Traditionally, a licensed nutritionist is needed to convert the nutrient level information into foodstuff level information as a meal plan. (Heinonen, 2009)

Nutri-Flow® Software gives personalized dietary guidance on the foodstuff level as foods and beverages by applying a national food composition database, national nutritional recommendations with personal health data and eating habits. Fuzzy logic handles the uncertainty and imprecise values, and the mapping from the nutrient level to foodstuff level is carried out with Fuzzy Expert System (FES), which contains licensed nutritionists' knowledge in a rule base. (Heinonen et al., 2009)

Optimization in Nutri-Flow® Software is needed to find a level of change for a set of foodstuffs to reach a more balanced diet. A type of an optimization problem usually defines which optimization methods are applicable within the problem domain. (Heinonen, 2009)

Direct search is one approach to solve optimization problems which have non-differentiable or discontinuous objective functions. Most traditional optimization methods require gradient or higher derivatives from the objective function to work properly. (Kolda et al., 2003)

Heuristic search is a set of optimization methods with rules to guide the optimization process towards to the global optimum. Genetic Algorithms (GAs) belongs to the heuristic search group. The theory was invented in 1975 and GA implements the concept from Darwinian principle of natural selection (Darwin, 1859). The terminology in GA is closely adapted from natural genetics (Holland, 1975). A solution is called a chromosome, which has locus bind variables. The solution can be coded in a binary form, but with many real world problems high precision makes chromosomes very long, and it gets inefficient. (Davis, 1991)

A set of solutions or a population is put in a competitive environment where each chromosome has a fitness value evaluated by an objective function (Holland, 1975), (DeJong, 1993). In every iteration round, GA operators are applied to the population. A crossover operator combines genetic materials of selected chromosomes creating a new population. Mutation operation makes random variation in the population to prevent the optimization process from stopping at local minimums. (Holland, 1975) Since the theory was published, several GA operators have been

developed with variations according to the current optimization problem (Goldberg, 1989a)

This paper presents the GA optimization model developed for the Nutri-Flow® Software.

2 Selecting an Optimization Method

The type of optimization problem is the key to selecting an optimization method. In this study, the optimization problem has a large amount of variables and there are several solutions which are almost equally good.

To form a mathematical function in this problem domain is challenging and time consuming. It is possible to calculate gradients at certain points, however the system is not continuous due to its nature. Based on this, traditional optimization methods are not selected for testing.

Direct search can be used with an optimization problem domain where is no gradient or higher derivative available. It works also with non-continuous problem space. (Kolda et al., 2003) In the optimization problem of this study, it might be possible to use direct search to find a single solution. This needs further testing.

A real-coded GA has solutions available as a set of comparable solutions. After evolving a population - the set of solutions - the set of the best options is available in the solution domain already without a mapping function. (Man et al., 1999)

The stochastic nature of GA with crossover, mutation, and elitism operators the optimization process is not easily stopped at local minimums. It is argued that real-coded GA does not always reach the best result. However, the real-coded GA is widely used with real word optimization problems. One example of this is represented in (Le and Kim, 2011).

There are also other Evolutionary Programming algorithms, e.g. differential evolution, which could be used in the optimization problem domain of this study.

3 Genetic Algorithms

Genetic algorithms are based on processing of a population of coded solution alternatives.

3.1 Chromosome Coding

A chromosome in GA represents a possible solution for the optimization problem. The size of search domain and accuracy level of result are used when evaluating if the coding should be done in the binary or real-value domain. (Goldberg, 1989a)

Real-value coding is used widely with practical real world optimization where the search domain is usually large and required high accuracy, where the binary coding would be inefficient with extremely long chromosomes. The solution domain is applied in real-value coding thus no result mapping is needed. It has been argued that the real-value coding does not always yield good results. (Man et al., 1999)

3.2 Population

A population is a term for a set of chromosomes. The best solution should be found by evolving the population by applying GA operators. In this process, the size of the population has an effect on convergence speed and reaching the global optimum. The population size can be fixed or it can vary throughout the optimization process. (Goldberg, 1989b)

Too small a population does not have enough diversity to evolve towards the global optimum. Longer the chromosome, bigger the population should be. (Goldberg, 1989b) When the population size is too big, the evolving needs more iteration rounds to reach the best solution. This affects the computing time with a slow convergence rate. (Affenzeller et al., 2007) There are statistical methods for generating the initial population, however it is usually generated randomly (Reeves and Rowe, 2002).

3.3 Crossover

Crossover operators evolve population towards better solutions by distributing good genetic matter between generations. Crossover starts with selecting parents by using a selection method, usually by the roulette wheel selection method (Sorsa et al., 2008) or by the tournament selection method (Goldberg, 1990).

Good genetic material is found into mating pool by selection methods since probability for selecting the fittest parents from the population is higher than for the worse ones. This is the basis how better solutions are found on every iteration round. (Sorsa et al., 2008)

The tournament selection method is based on randomly selected chromosomes from the population, and the chromosome with the best fitness value is selected. The tournament size k defines how many chromosomes will be selected from the population. Typically, value for it is 2. (Goldberg, 1990)

After a mating population is formed by a selection method, a crossover operator is applied. A design parameter for the crossover operator is crossover probability p_c, which determines if the current parent chromosomes are combined with the crossover operator or if they are moved directly to the offspring population. (Man et al., 1999)

With real-value coded chromosomes, uniform and non-uniform crossover operators can be applied. Where uniform operators act in the similar way in every generation, and non-uniform operators work depending on the age of the population. Two offspring from two parents are formed by an arithmetic operator as follows:

$$y_i^1 = \alpha_i x_i^1 + (1-\alpha_i) x_i^2, \quad (1)$$

$$y_i^1 = \alpha_i x_i^2 + (1-\alpha_i)x_i^1,$$

where α_i are uniform random numbers. α can vary in non-uniform crossovers, and is constant in uniform crossovers. (Sorsa et al., 2008). New crossover operators are developed actively (Gegúndez et al., 2007).

3.4 Mutation

Mutation is a GA operator which creates a random variation in the population. It maintains the population diversity by generating new genetic material. With correct design parameters, mutation prevents GA from stopping at local minimums. (Sorsa et al., 2008)

Uniform and non-uniform mutation operators can be used with real-coded chromosomes. The uniform mutation operator can be applied for each gene in a chromosome using the same mutation probability where initial population is $x_i^t = \langle v_1, \ldots, v_n \rangle$, after a mutation operation with 1≤n it becomes $x_m^t = \langle v_1, \ldots, v'_k, \ldots, v_n \rangle$. The random value for v'k is in a feasible range for the locus k. (Michalewicz, 1996)

One of the most commonly used non-uniform mutation operator is the Michalewicz's non-uniform mutation (Michalewicz, 1996), where the mutated element v'k is calculated by

$$v'_k = \begin{cases} v_k + \Delta(t, u_k - v_k), \text{if a random digit is 0} \\ v_k + \Delta(t, v_k - l_k), \text{if a random digit is 1} \end{cases}, \quad (2)$$

where

$$\Delta(t, y) = y\left(1 - r^{\left(1-\frac{t}{T}\right)^b}\right). \quad (3)$$

Mutation probability p_m controls the mutation operations. When using too big values for pm, good genetic material can be lost and the convergence rate slows down. With very low values, there is little or no effect on the population.

3.5 Elitism

When altering the genetic pool of a population with GA operators, there is a possibility to lose the best solution of an iteration round. The elitism operator is designed to save the best chromosome and transfer it to the next generation. This is done usually by replacing the worst chromosome from the new population with the best chromosome from previous population. (Man et al., 1999)

4 Genetic Algorithms in Domain of Dietary Optimization

In this study, the GA optimization is an interconnected module in Nutri-Flow® Software as represented in Figure 1. Nutri-Flow® Software has a database where personal dietary habits and personal health data are stored when filled in. A nutritional state of a diet is assessed and the guidance is mapped to foodstuff level, e.g. as foods and beverages. (Heinonen et al., 2009)

FES has two hierarchical levels representing main and sub-groups in the classification of foodstuffs in Nutri-Flow® Software database. The input consists of 30 nutrient variables and the output of FES has 129 variables, including 87 sub-groups of foodstuff classification and 42 foodstuff variables. The rule base in FES has 590 rules on the first hierarchical level, and 400 rules on the second level. (Heinonen, 2009) The expertise is introduced with the rules. Development of FES is presented more detailed in (Heinonen, 2009; Heinonen et al., 2009).

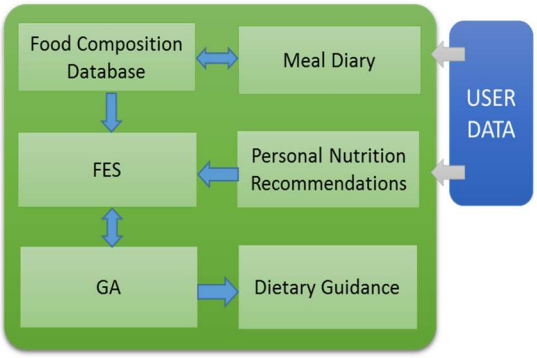


Figure 1. Schematic model of Nutri-Flow® software (Heinonen, 2009).

The output of FES is used to form a search space for GA. An initial population is generated randomly using the search space. A gene in a chromosome represents a foodstuff with recommended daily intake level. A chromosome represents a dietary recommendation in foods and beverages with their recommended daily intake levels.

An objective function is formed to analyze the fitness of the chromosomes. The recommendation takes into account also personal taste and allergies with the nutritional status.

A diagram of the GA optimization process is presented in Figure 2. Because the initial population is a random set, it is recommended to run GA optimization several times with a new initial population. GA in the Nutri-Flow® software is set to run 10 re-runs with a new initial population each time. All the best solutions of the last population of each GA re-run are stored, and the best of the best is used to form the dietary guidance. The selection of GA re-runs and other GA design parameters are described later on this paper.

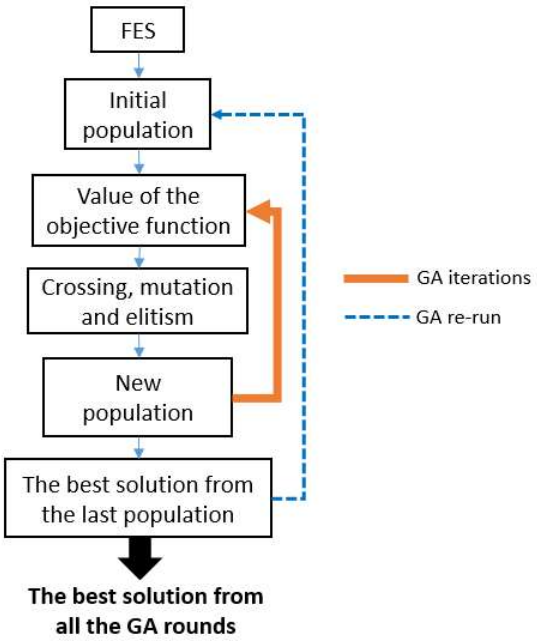


Figure 2. The GA Optimization process.

5 Developed Model

5.1 Search Space

The search space is formed according to the output of FES. The output stores the information about the foods to be added or reduced with the importance of the change. There are three different sets of foodstuffs in the output: "reduce", "no change" and "add".

There is a need to assess feasible ranges of daily intake levels for the foods, because the initial population is generated randomly within the given intake range for each food. The intake range for foods to be reduced is formed as [0, *i*], where *i* is the initial intake level. The foods, which have no need to be changed, will be kept at the original intake level, *i*. Foods to be added have a range with [*i, m*], where *m* represents a recommended maximum intake level, defined by licensed nutritionists. If no preset for maximum intake is present, the maximum value is evaluated from the initial intake level.

The search space has an effect on performance and calculation time of Nutri-Flow® software. If the initial population is biased due to infeasible range of intake levels of foods, is the convergence of the fitness value slow, and the global optimum might never be reached.

5.2 Objective Function

In this study, the objective is to minimize the fitness value. The objective function uses FES to determine the nutritional status of a solution. FES fuzzifies the nutrient and gives membership grades for each fuzzy membership function. The goal is to minimize "too little" and "too much" membership grades μ. The FES membership functions are represented in Figure 3.

Personal nutritional recommendations for 30 nutrients are used to tune the membership functions, data points A, B and C presented in Figure 3. The data points A, B and C represent Lower Intake level (LI), Recommended Intake level (RI), and Upper Intake level (UL), respectively (Heinonen, 2009).

There is no need to use normalizing function for each nutrient input levels since membership grades are already normalized to [0,1].

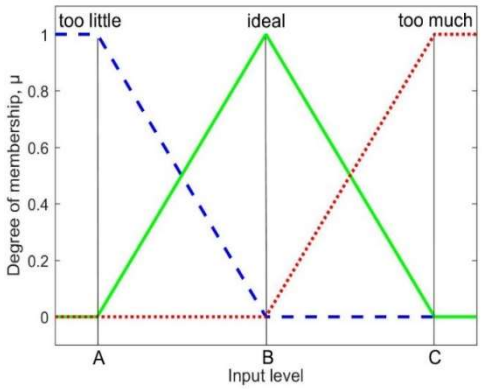


Figure 3. FES Input Membership Functions (Heinonen, 2009).

There are also other objectives to be minimized which are represented in the objective function as follows:

$$MIN(a\sum_{1}^{n}(l_n\mu_{l,n}+u_n\mu_{u,n})+b\sum_{1}^{m}|d_m|+c\sum_{1}^{p}k_p), \quad (4)$$

where a is weighting coefficient for nutrition status
l_n is weighting coefficient for importance of nutrient n at "too little" condition,
$\mu_{l,n}$ is membership grade for "too little" membership function for nutrient n,
u_n is weighting coefficient for importance of nutrient n at "too much" condition,
$\mu_{u,n}$ is membership grade for "too much" membership function for nutrient n,
b is weighting coefficient for level of change
d_m is distance from current diet for foodstuff m
c is weighting coefficient for importance level of other variables
k_p is other diet related variables.

The first term represents the difference between the recommended nutrient intake level and the current intake level in the fuzzy domain. According to licensed nutritionists, some nutrients are more important in keeping close to the recommendation level than others. The weighting factors l_n and u_n are used to emphasize the importance of the nutrient according to intake level. The weighting factors are tuned by licensed nutritionists. In this study, 30 nutrients are taken into account when assessing the individual nutritional state.

Smaller steps are usually easier to follow when changing eating habits. The second term is a measure how much a chromosome would change the initial diet.

It measures an amount of foodstuffs to be changed and the magnitude of the change. The weighting factor b is used to tune the level of change, and can be altered by a user through the user interface. If the value is set to 0, the magnitude of change is not taken into account.

The third term contains other measurable variables needed in dietary guidance, such as vegetable intake level and energy intake level. Other non-nutritional variables can be used too when these are used to fine-tune the recommendation. These should not have an advert effect on the nutritional status of the diet.

5.3 Coding

It was found out already in (Heinonen, 2009) that binary coding is not a good option to perform GA in this problem and solution domain, therefore the real-valued coding was selected. With real-valued coding, a gene stores a name of a foodstuff with a daily intake level: "a slice of rye bread", "20 g". The length of a chromosome depends on the variation of food beverage items in the meal diary during the period to be evaluated, therefore the length of the chromosomes is not fixed.

According to licensed nutritionists, with a recommended three to seven day period of meal tracking, there are average 20 different food items. Similar foods are combined when assessing daily intake levels. To confirm this, there is a need to do statistical evaluation on the meal diary database.

5.4 Population Size

While the chromosome size is not constant, the population size can vary throughout the different optimization run. In this study, the population size is kept constant at 100.

Population size has a strong effect on computing time when the size is too big as shown in (Heinonen, 2009). The effect on varying population size can be monitored through computing time, convergence rate, and the final result.

5.5 Crossover

Arithmetic crossover and tournament selection are used in this study. The design parameters for tournament selection are crossover probability p_c and tournament population size p_s. The selected arithmetic crossover is uniform.

Different values for p_c and p_s are tested while other GA design parameters kept fixed. The performance of the overall system is monitored via test parameters such as the convergence rate of the best and average solution, and fitness of the best solution.

5.6 Mutation and Elitism

The mutation and elitism operators are applied in this study to prevent the optimization process from stopping any local minimum. Elitism is used to prevent of loss the best solution during evolving the population.

Mutation probability m_p is evaluated with different value ranges while other parameters kept fixed. The overall system is monitored with the same parameters as crossover, the convergence rate of the best and average solution, and the fitness of the best solution.

Elitism is set to keep the best solution of the current population and replace the worst solution from the offspring population. In this study, the elitism operator replaces only one chromosome in the offspring population.

5.7 Additional Parameters

Each diet and goal are different when assessing personal dietary guidance. There is no way to set a global fitness value when the optimization process should be exited. The guidance should be able to follow without too drastic changes in the diet. Therefore, there are additional parameters in the objective function to make the guidance easier to follow.

There should be a certain exit point for the optimization process, which is executed using a counter for iterations. Too small a value for the counter would stop the optimization process too early with a poor result, and a too large number would just waste computing time. The optimum value for iterations can be found when the main design parameters are tuned first. The optimization process can be stopped when the best possible solution is not changing with a certain amount of iterations, but the average fitness value still evolves. In this study, the exit point for iterations is a fixed number, but in the future, the algorithm could analyze the convergence speed of the best solution vs. average solution and stop it when evolving for the best solution is not detected.

The initial population of GA is random. There is a possibility that the solution set has very bad fitness values or has too low a variation between the solutions. This could prevent GA to find the global optimum and cause a stop at a local minimum. This can be prevented to re-run GA with new initial population as presented in Figure 2. The rate of GA re-run should be set high enough. This multiplies the amount of iterations, thus has a direct effect on the computing time. A fixed value for the re-run rate is used in Nutri-Flow® software. The testing of the effect of rate on the re-runs requires a big set of meal diaries, hence it will be done later.

5.8 Java Genetic Algorithm Package

Nutri-Flow® software is currently written in Java and it applies JGAP – Java Genetic Algorithm Package library which provides the GA operators with all necessary design parameters.

Validating Nutri-Flow® software is done with data sets run in Nutri-Flow®, and in Matlab® model. The Nutri-Flow® software is working correctly if the same input generates the same output. GA needs a special focus, since optimization results with one input could

create several different outputs. Generally, there are several different ways to alter initial diet to achieve a certain balanced diet. Validation of GA module requires a large set of test runs with the statistical and nutritional analysis of the result set.

5.9 Data Acquisition

To keep the results comparable throughout the development work of Nutri-Flow® software, there is a set of individuals and data of their dietary habits. The test data represents different average dietary habits from fast food diet to vegan diet. The set of individuals contains male and female persons with different health profiles from athlete to pregnant women, and slim to fat inactive persons.

The used test data is suitable for testing GA optimization performance on the current problem domain. Data acquisition is done using Nutri-Flow® software. All the test variables of GA optimization process are saved to a text file and all the numerical data is analyzed with Excel or Matlab® which also provides a good platform for testing new calculation ideas.

6 Results and Discussion

6.1 Overall Performance of Nutri-Flow® software

Nutri-Flow® software provides the output as dietary guidance. Also the performance of GA is recorded separately to provide numerical data for analysis. The overall performance of the system is analyzed by licensed nutritionists.

When a person applies the recommended actions to the initial diet, the result must lead to a better state of nutrition. This can be analyzed numerically by comparing personal recommendations with the nutritional state of the recommended diet. The comparison should be done also in the fuzzy domain, since the values include uncertainty and imprecision. The formed dietary recommendation should be also feasible with recommended foods and their portion sizes and with level of change. For example, nobody would like to eat three tablespoons of cinnamon.

Results of the overall performance are promising. All the requirements are met numerically evaluated. However some of the recommended changes in diet were controversial, since there are expectations that some foods are healthier than others. Some healthy considered foods were recommended to reduce while some other foods were recommended to increase. After a review of licensed nutritionists, also the controversial guidance has been accepted, since the healthiness of a single foodstuff depends on the individual overall diet, not on the health claims on single food.

The tests revealed that, there are cases when GA is not working. Optimization with very limited diets with strong dietary limitations did not lead to any result. The reason is that, there was not enough variation in the population to find a feasible result. Other case was with very balanced diets. The initial diet was already so close to the recommendations that GA did not always find any better solution.

A statistical approach with a large set of individuals and meal diaries is needed to get more comprehensive data to analyze the performance of the overall system generally.

6.2 Population Size

According to previous contextual testing, the population size was set a fixed value, 100 individuals. There is a need to carry out more tests to show that selected population size is adequate for varying chromosome size. This could be done with the planned statistical testing for the overall system.

However, a population size more than 1000 will lead to longer computational time, which might affect the usability of the service. With current test data, global optimum is reached with current design parameters with all test cases.

6.3 Crossover

In this study, only uniform arithmetic crossover with the tournament selection method was analyzed. Parameters crossover probability p_c and tournament population size p_s are tested with population size 100. The best working values for the parameters are 0.8 and 2 for p_c and p_s, respectively. The test was run with the same data set as with the population test.

6.4 Mutation and Elitism

The same initial test data was applied also with testing mutation and elitism. The population size with this test was kept fixed at 100, and p_c and p_s to 0.8 and 2, respectively. Too big a value for mutation probability m_p slows the convergence rate as it affects the better genetic material more probably. The value 0.02 was selected. According to the data, the best value for m_p was 0.02.

Test runs show that running GA operators without elitism will lead us to lose some of the best solutions during the evolving of the population. This can be seen in Figure 4 where the trend line has peaks for both, the average fitness value and for the best fitness value.

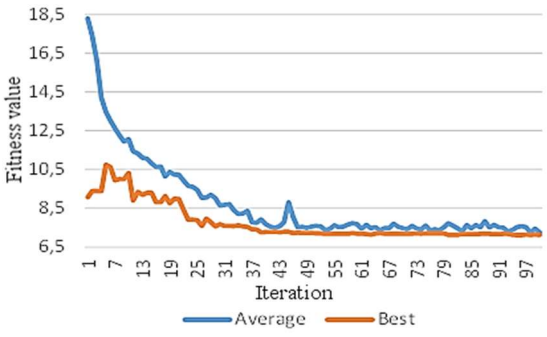


Figure 4. GA optimization without elitism operator.

6.5 Additional Parameters

According to the overall assessment with the numerical data of iterations, the best solution was reached with most of the test data within 100 iteration rounds. It is possible to create a self-monitoring feature to stop iterations when no better solutions are not found after a certain amount of iterations.

The initial population is created with random values within the given range. Sometimes it leads to a bad population which does not lead to the global optimum. GA should be run several times to find the global optimum. In Nutri-Flow® Software GA is run 10 times at the moment. This needs also further testing with a larger test set.

7 Conclusions

The nutritional guidance tool combines expertise with extensive food composition data through optimization. According to the overall assessment, the results are promising. All the requirements were met except two cases where no result was found at all. The real-coded chromosomes with GA operators such as crossover, mutation, and elitism can be used in the domain of dietary guidance when the search space is formed by FES. The objective function is crucial in comparing the results. When nutrient intake levels are handled in the Fuzzy domain, the imprecision and uncertainty can be taken into account, too.

The feasible result and a short computation time are essentials to make the Nutri-Flow® software usable. The validation of the system was carried out with expert knowledge, comparisons of nutritional status, and monitoring the key features of GA performance. There is a need to carry out more intensive testing with a large test data set. Also other optimization methods suitable for this problem domain will be tested.

References

M. Affenzeller, S. Wagner and S. Winkler. Self-Adaptive Population Size Adjustment for Genetic Algorithms, *EUROCAST 2007, LNCS 4739*, pages 820–828, 2007. doi: 10.1007/978-3-540-75867-9_103.

C. Darwin. *On the origins of species by means of natural selection*. Murray, London. 1859.

L. Davis. *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York. 1991. ISBN 0-442-00173-8.

K. DeJong. *Genetic algorithms are NOT function optimizers*, Whitley LD (ed) FOGA – 2, Morgan Kaufmann, Los Altos, CA, pp 5–17, 1993.

M. E. Gegúndez, P. Palacios and J. Álvarez. A New Self-adaptative Crossover Operator for Real-Coded Evolutionary Algorithms, *ICANNGA 2007, Part I, LNCS 4431*, pages 39–48, 2007. doi: 10.1007/978-3-540-71618-1_5.

D. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wiley Publishing Company, Massachusetts. 1989a.

D. E. Goldberg. Sizing Populations for Serial and Parallel Genetic Algorithms, In: *Proc. 3rd International Conference on Genetic Algorithms*, pages 70–79, 1989b.

D. Goldberg. *A note on Boltzmann tournament selection for genetic algorithms and population-oriented simulated annealing*, Tech. Rep. Nb. 90003, Department of Engineering Mechanics, University of Alabama, Tuscaloosa, A. 1990.

P. Heinonen. *Fuzzy Expert System and GA optimization in Dietary Guidance Application*, Master's Thesis, University of Oulu, March 2009.

P. Heinonen, M. Mannelin, H. Iskala, A. Sorsa and E. Juuso. Development of a Fuzzy Expert System for a Nutritional Guidance Application, *Proc. Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, Lisbon, Portugal, July 20-24, 2009*, pages 1685-1690, 2009. ISBN 978-989-95079-6-8.

J. H. Holland. *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*, University of Michigan Press. 1975.

IFPRI. *Global Nutrition Report 2016: From Promise to Impact: Ending Malnutrition by 2030*, International Food Policy Research Institute, Washington, DC. 2016.

T. G. Kolda, R. M. Lewis and V. Torczon. Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods, *SIAM Review*, 45(3): 385-482, 2003. doi:10.1137/S0036144502428893.

T.-H. Le and D.-J. Kim. Application of a real-coded genetic algorithm for the fitting of a ship hull surface through a single non-uniform B-spline surface, *Journal of Marine Science & Technology*, 16(2): 226-239, 2011. doi: 10.1007/s00773-011-0118-1.

K. F. Man, K. S. Tang and S. Kwong. *Genetic algorithms: concepts and designs*, Springer, Springer. 1999. doi: 10.1007/978-1-4471-0577-0.

Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs*, 3rd ed., Springer-Verlag, Berlin Heidelberg New York, 1996. ISBN 3-540-58090-5.

C. R. Reeves and J. E. Rowe. *Genetic Algorithms - Principles and Perspectives: A Guide to GA Theory*, Kluwer Academic Publishers, Boston. 2002. doi: 10.1007/b101880.

A. Sorsa, R. Peltokangas and K. Leiviskä. Real-coded Genetic Algorithms and Nonlinear Parameter Identification, *IEEE International Conference on Intelligent Systems*, Varna, Bulgaria, September 6-8, 2008.

Modular Model Predictive Control Concept for Building Energy Supply Systems: Simulation Results for a Large Office Building

Barbara Mayer[1] Michaela Killian[2] Martin Kozek[2]

[1]Institute of Industrial Management,FH Joanneum, Austria, {barbara mayer}@fh-joanneum.at
[2]Institute of Mechanics and Mechatronics, Vienna University of Technology, Austria, {michaela killian, martin kozek}@tuwien.ac.at

Abstract

The management of modern large buildings' energy supply systems (ESS) demands the maximal usage of renewable energy sources at minimum energy costs while meeting the energy demand of the consumption zone. Building ESS for heating and cooling usually consist of various supply circuits with several energy sources and different physical characteristics, possibly incorporating switching aggregates (heat pump, chiller) with latency times and stratified storage which change their operating state in a discontinuous fashion. Hence, these circuits can be seen as hybrid systems whose modelling as well as optimisation are demanding. Model predictive controllers (MPC) are an effective means for the optimisation of such problem formulations with divergent goals. The proposed modular predictive control concept (MPCC) is designed for a flexible usage in ESS building automation adding one appropriate MPC for each supply circuit including mixed-integer MPCs to the individual building's control structure. The resulting MPCC is capable of maximising the usage of renewable energy sources at minimum cost as well as efficiently managing switching aggregates with active storage. Suitable modelling of the linear and hybrid systems is demonstrated and validated on the example of a large office building in Austria. Furthermore, a simulation study shows the performance of the resulting MPC concept compared to a rule-based controller.

Keywords: building energy supply system, model predictive control

1 Introduction

Building control has become an important field since the building sector is responsible for about 40% of the total energy consumption in developed countries, (Pérez-Lombard et al., 2008). Therefore, an increasing effort has been put both on the development of energy saving building physical structure such as passive heating and cooling systems and on energy efficient building operation techniques. Model predictive control (MPC) has been proven as a promising technology for such building systems in recent years, (Širokỳ et al., 2011), with the ability of incorporating most influential quantities such as weather forecasts, (Oldewurtel et al., 2012), or occupancy information as well as technical constraints into the prediction. Moreover, the controller's optimisation target can include conflicting objectives expressed in thermal comfort and economic trade off. However, modelling of building systems is a crucial part for predictive building control, (Privara et al., 2013).

The efficient operation of large buildings' energy supply systems (ESS) aims at a maximal usage of renewable energy sources at minimum energy costs. Energy supply systems (ESS) of large buildings usually consist of various supply circuits of different physical characteristics, including heat exchangers and continuous pumps or additionally switching aggregates such as chillers or heat pumps with successive active energy storage, e.g. stratified storage tanks. Obtaining accurate models of latter systems is difficult, (Liu and Henze, 2004), since stratified storage is operated in either charging or discharging mode which is directly influenced by the state of the aggregate and its limitation of minimal up and down times. Due to the mixture of continuous and discrete variables hybrid models are suitable to approximate those systems. The efficient control of hybrid systems is challenging. For the optimisation within a predicitve controller with quadratic target this leads to a constrained mixed-integer quadratic problem (MIQP) at each time step. This can either be solved by a dual stage procedure where the storage tank operation mode profile is firstly fixed and the remaining QP solved in a second step, (Ma et al., 2009), or in a single step e.g. by using a branch and bound algorithm, (Mayer et al., 2016).

The classic approach to controlling buildings, especially ESS, is rule based control (RBC) due to its simplicity. However, RBC does not allow advanced management of e.g. active storage. Advanced control approaches of the recent years propose one single MPC controlling the entire building comprising the buildings' zone control as well as the energy supply optimisation, (Oldewurtel et al., 2012; Privara et al., 2013). While the operation with one global MPC would guarantee optimality, it is too rigid for application in industrial building automation.

This work introduces a modular predictive control concept (MPCC) for the energy supply level using dedicated predictive controllers for the respective circuits' physical characteristics. Basic circuits with heat

exchangers and continuous pumps can be approximated by linear state space systems, whereas the hybrid models are represented by piecewise affine (PWA) systems. All models are analytically derived based on thermodynamic principles and energy balance equations, (Berkenkamp and Gwerder, 2014; Mayer et al., 2015). The overall structure is capable of optimising heating and cooling supply at minimum cost and maximal usage of renewable energy sources. Furthermore, an efficient management of the active storage connected with the switching aggregate is guaranteed due to the application of a dedicated mixed-integer MPC (MI-MPC). The MPCC is designed to be embedded into a hierarchical building control structure, decoupling the energy consuming office level (OL) from the energy supply system level. In the OL a predictive controller is assumed to maximise user comfort while minimising the supply energy respecting disturbances, e.g. (Killian et al., 2015). It thus provides the reference trajectory for heating and cooling power for the MPCC. The modularity of the MPCC enables a flexible application in industrial building automation using dedicated models and MPCs allowing separate implementation and tuning. The main contributions of the paper therefore are:

- A modular predictive control concept (MPCC) for building energy supply systems (ESS) maximising the usage of renewable energy sources at minimum cost.
- Supply circuits with switching aggregates and stratified storage are modelled as hybrid systems and controlled by MI-MPCs aiming at an efficient management of aggregates and storage.
- Flexible application in industrial building automation leading to an application specific control architecture.
- A simulation study showing the performance of the proposed MPCC compared to a rule-based controller applied to a demonstration building in Austria.

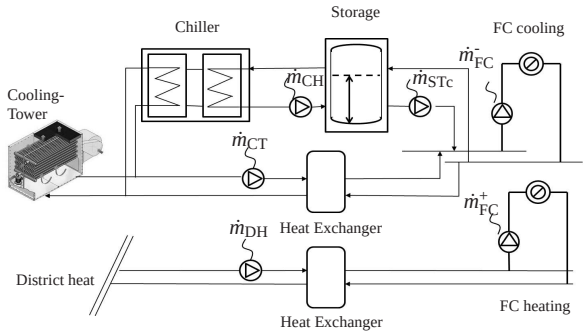
The remainder of this paper is structured as follows: In Section 2 the demonstration building setup as well as the overall control structure is given. The building models for the ESS dedicated for the different MPCs are explained in Section 3, whereas the predictive controllers are introduced in Section 4. The simulation results for the demonstration building are shown in Section 5 and finally, a conclusion is drawn in Section 6.

2 Modular Model Predictive Control Concept

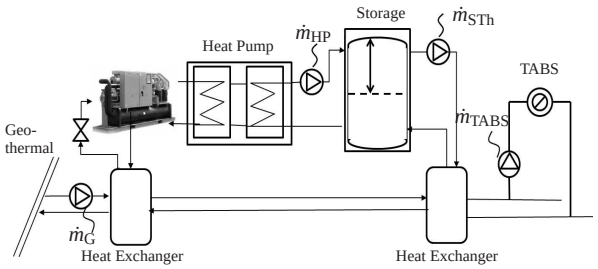
Within this Section the building setup and the overall MPCC for the demonstration building is given.

2.1 Building Setup

The building which is presented in this work is a $27.000m^2$ University building in the center of Salzburg, Austria. It has five floors above ground containing several large and numerous smaller meeting rooms, offices and lecture rooms. For this work focus is put on the second and third floor, which comprise of about 500 rooms of some $13.000m^2$, almost all used as offices. Both floors are supplied by Fan Coils (FC) and a Thermal Activated Buildings System (TABS). The ESS of this building consists of heating and cooling supply circuits for FC and TABS, see Figure 1. The FC supply is split into cooling supply based



(a) FC circuits including free cooling, the chiller, and district heat.



(b) TABS circuits including geothermal source and the heat pump.

Figure 1. Energy supply circuits for cooling and heating of the University building in Salzburg, Austria.

on free cooling and chiller circuit and heating supply from the district heat. TABS has only one piping system supplied by the geothermal source and is routed via the heat pump circuit in case of heating.

2.2 Modular Predictive Control Concept

The modular predictive control concept (MPCC) generally consists of independently acting MPCs with the same target - one MPC for each supply circuit comprising the corresponding energy source as well as the supply system of the building. Basic circuits consisting of heat exchangers and continuous pumps can be approximated by linear models. Consequently, linear MPCs (LMPC) are applicable in this case. However, if switching aggregates have to be considered, the predicitve controller not only has to optimise the continuous manipulated variables but also the discrete aggregate's switching state. Moreover, the coupled active storage changes its operating mode in

a discontinuous fashion depending on the switching instance. Therefore, a dedicated mixed-integer MPC (MI-MPC) is applied optimising latter circuits considered as hybrid systems. The control structure for the proposed ESS, see Figure 2, is designed according to the available energy sources and aggregates shown in Figure 1(a) and Figure 1(b) which distinguishes between FC and TABS heating and cooling supply. For the FC cooling system an LMPC controls the free cooling circuit, whereas an MI-MPC manages the cooling supply from the chiller including a stratified storage tank. The FC heating supply is sourced by the district heat and controlled by the second LMPC. The TABS is supplied by the geothermal source either used directly in the case of cooling or routed via the heat pump and subsequent stratified storage tank if heating is required. Hence, an LMPC is active for cooling and an MI-MPC for heating. Furthermore, the ESS control structure is embedded into an overall hierarchic building control structure with a predicitve controller for the OL, e.g. (Killian et al., 2015), which is responsible for providing a prediction of the required heating and cooling power, $\dot{Q}_{\mathrm{FC}}^{\mathrm{ref,+}}$ respectively $\dot{Q}_{\mathrm{FC}}^{\mathrm{ref,-}}$, for each energy supply system. All variables and parameters used are listed in Table 1 and the two types of MPCs are formally described in Section 4.

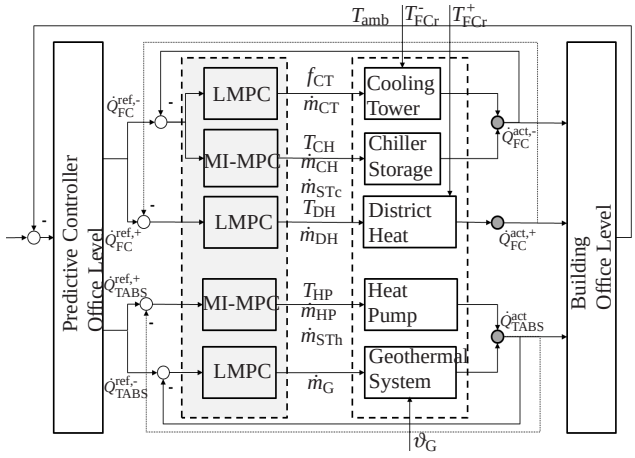


Figure 2. Control structure for cooling and heating circuits for FC and TABS embedded into a hierarchical building control structure.

Table 1. Nomenclature.

Variable	Description	Unit
$\dot{m}_{\mathrm{CT}}$	mass flow from cooling tower	[kg/s]
f_{CT}	fan speed of cooling tower	[m/s]
T_{amb}	ambient temperature	[°C]
T_{FCr}^{-}	return temperature for FC cooling	[°C]
T_{CH}	supply temperature of chiller	[°C]
$\dot{m}_{\mathrm{CH}}$	mass flow from chiller to storage	[kg/s]
$\dot{m}_{\mathrm{STc}}$	mass flow from storage of cooling circuit	[kg/s]
T_{DH}	supply temperature of district heat	[°C]
$\dot{m}_{\mathrm{DH}}$	mass flow from district heat	[kg/s]
T_{FCr}^{+}	return temperature for FC heating	[°C]
T_{HP}	supply temperature of heat pump	[°C]
$\dot{m}_{\mathrm{HP}}$	mass flow from heat pump to storage	[kg/s]
$\dot{m}_{\mathrm{STh}}$	mass flow from storage of heating circuit	[kg/s]
$\dot{m}_{\mathrm{G}}$	mass flow from geothermal source	[kg/s]
ϑ_{G}	difference geothermal supply, return temp.	[°C]
$\dot{Q}_j^{\mathrm{ref,+}}$	reference heat flow for supply j heating	[kW]
$\dot{Q}_j^{\mathrm{ref,-}}$	reference heat flow for supply j cooling	[kW]
$\dot{Q}_j^{\mathrm{act,+}}$	actual heat flow for supply j heating	[kW]
$\dot{Q}_j^{\mathrm{act,-}}$	actual heat flow for supply j cooling	[kW]
j	supply: TABS or FC	

3 Energy Supply Level - ESS - Models

In this Section the modelling strategy and structure for the circuits of the ESS shown in Figure 1 is given. Each circuit is modelled individually.

3.1 Linear Models

For building control aspects the achievable heating or cooling power, $\dot{Q}$, of each supply circuit is relevant, which can be approximated by simplified energy balance equations: $\dot{Q} = \dot{m} \cdot \Delta T \cdot cp$, where $\dot{m}$ denotes the mass flow, ΔT the difference of supply and return water temperature, and cp the specific heat capacity of water. Heat exchangers and continuous pumps preferably work around a certain operating point. For the geothermal and the district heat supply two linearised static models based on those thermodynamic principles are analytically derived. The geothermal model contains one manipulated variable, $\dot{m}_{\mathrm{G}}$, one disturbance, ϑ_{G}, and the output $\dot{Q}_{\mathrm{TABS}}^{-}$. The Δ-variables denote the deviation of the variables to the operating point and $\mathrm{COP_G}$ the coefficient of the geothermal performance:

$$\Delta\dot{Q}_{\mathrm{TABS}}^{-} = \underbrace{\mathrm{COP_G} \cdot \mathrm{cp}}_{const.} \cdot [\dot{m}_{\mathrm{G}}|_o \cdot \Delta\vartheta_{\mathrm{G}} + \vartheta_{\mathrm{G}}|_o \cdot \Delta\dot{m}_{\mathrm{G}}]. \quad (1)$$

The district heat model is derived in the same way and contains two manipulated variables, T_{DH} and $\dot{m}_{\mathrm{DH}}$, one disturbance, T_{FCr}^{+}, and the output $\dot{Q}_{\mathrm{FC}}^{+}$:

$$\begin{aligned}\Delta\dot{Q}_{\mathrm{FC}}^{+} \quad &= \underbrace{\mathrm{COP_{DH}} \cdot \mathrm{cp}}_{const.} \cdot [\dot{m}_{\mathrm{DH}}|_o \cdot (\Delta T_{\mathrm{DH}} - \Delta T_{\mathrm{FCr}}^{+}) \\ &\quad + (T_{\mathrm{DH}}|_o - T_{\mathrm{FCr}}^{+}|_o) \cdot \Delta\dot{m}_{\mathrm{DH}}]. \end{aligned} \quad (2)$$

As depicted in Figure 1(a) free cooling is based on the cooling tower. Free cooling is exclusively used if the chiller is inactive. The ambient temperature, T_{amb}, constrains the cooling tower's operation for free cooling which is also a main disturbance next to the return water temperature T_{FCr}^{-}. Modelling cooling towers analytically aims at detailed complex models with non-linear dynamics of high order which are not suitable for the usage within MPCs. Hence, black-box identification is expedient if

historic data of the cooling tower in operation is available. Since the cooling tower can be operated at two fan speeds $i \in \{1,2\}$, the system can be approximated by two linear static models:

$$\dot{Q}_{\mathrm{FC}}(i) = c_{i,1} \cdot T_{\mathrm{amb}} + c_{i,2} \cdot T_{\mathrm{FCr}} + c_{i,3} \cdot \dot{m}_{\mathrm{CT}} + c_{i,4}, \quad (3)$$

where the corresponding coefficients $c_{i,k}$ have to be estimated within the black-box identification routine using historic data of the free cooling system. Note that the model in Eq. (3) is linear in the parameters $c_{i,j}$, thus least squares methods can be employed for optimal parameter estimation, (Ljung, 1999).

3.2 Hybrid Models

The chiller and the heat pump are switching aggregates with latency times, such that minimum up and down times have to be respected, e.g. after switching on the aggregate it must operate for a minimum up time until it can be shut off again and vice versa. Furthermore, the stratified storage tanks can operate in two basic modes: charging and discharging. Each mode is represented by a dedicated model with continuous variables. The discrete switching state of the aggregate (on/off), l_{on}, specifies which mode is exclusively active at each time. These operation modes further depend on the difference of the mass flows to, $\dot{m}_l$, and from, $\dot{m}_{\mathrm{STk}}$, the storage tank where $l = \{\mathrm{CH},\mathrm{HP}\}$ denotes the aggregate and $k = \{\mathrm{c},\mathrm{h}\}$ the cooling or heating circuit, (Mayer et al., 2015). The hybrid dynamic models contain three continuous states, $x(t)$, namely the thermocline of the storage $z_k(t)$, the temperature of the cold respective hot water $T_k(t)$, and the cooling respective heating power $\dot{Q}_j(t)$. The three continuous manipulated variables, $u(t)$, are the supply temperature of the aggregate $T_l(t)$, the mass flow from the aggregate to the storage $\dot{m}_l(t)$, and the mass flow from the storage to the TABS or FC supply $\dot{m}_{\mathrm{STk}}(t)$. $l_{\mathrm{on}}(t)$ is the discrete manipulated variable representing the switching state of the corresponding aggregate l for each time t. The continuous output $y(t) = \dot{Q}_j(t)$ denotes the cooling respective heating power for the supply j to the building. For each operating mode a model is firstly analytically derived through first order differential equations and secondly linearised at a fixed operating point. The non-linear dynamics can then be approximated by a hybrid system state-space formulation with discrete as well as continuous inputs represented by a piecewise affine (PWA) system:

$$x(t+1) = A_h x(t) + B_h u(t), \qquad \text{if } \delta_h(t) = 1, \quad (4a)$$
$$y(t+1) = Cx(t), \quad (4b)$$

where the binary variables $\delta_h(t) \in \{0,1\}, \forall h = 1,..,3$, are introduced to denote the status of the operation mode h and the system matrices are $A_h \in \mathbb{R}^{3\times3}$ and $B_h \in \mathbb{R}^{3\times4}$. $C = \begin{bmatrix}0 & 0 & 1\end{bmatrix}$ for all three modes since the third state is equal to the output. The explicit matrices for the chiller circuit model are given in (Mayer et al., 2016), whereas for the modelling of the heat pump circuit the reader is referred to (Mayer et al., 2015).

4 Model Predictive Controllers

The formal description of the two MPC types used within the MPCC and their common objective are given in this Section.

4.1 Objective Function

The predictive controller for the OL, e.g. (Killian et al., 2015), is responsible to guarantee user comfort by keeping the indoor temperature within a certain tolerance at minimum energy demand. The objective for the ESS on the other hand is to minimise the deviation to the required power and the energy costs.

For all MPC implementations a quadratic optimisation target is used, where the deviation to the reference trajectory of the cooling respective heating power is penalised with factor Q_y. Furthermore, the energy costs represented by Q_u caused by the manipulated variables are taken into account. Both additive terms are considered for each time step t over the whole prediction horizon Np. The Δ-variables again denote the deviation from the corresponding operating point. The formal description of the objective is given by:

$$J^\star = \min_{\Delta u \in U} \sum_{t=0}^{Np-1} \left\| \Delta\dot{Q}_j^{\mathrm{ref}} - \Delta\dot{Q}_j^{\mathrm{act}} \right\|_{Q_y}^2 + \left\| \Delta u(t) \right\|_{Q_u}^2, \quad (5)$$

4.2 LMPC

The three LMPCs for this work have a quadratic optimisation target as presented in Eq. (5). They only differ in the model they rely on. The FC model for free cooling uses a linear model with absolute inputs, whereas the TABS geothermal model and the FC district heat model are a linearised model with Δ- variables which denote the deviation from the corresponding operating point. All three LMPCs are implemented according to the receding horizon strategy, (Camacho and Alba, 2013).

4.3 MI-MPC

In Section 3 the hybrid models for the chiller and the heat pump circuits are motivated. The objective is given in Section 4.1. The corresponding constraints are:

PWA model as given in Eq. (4),
$$\delta_h(t) \in \{0,1\}, \quad (6a)$$
$$\sum_h \delta_h(t) = 1, \quad (6b)$$
$$u_{\min} \le u(t) \le u_{\max}, \quad (6c)$$
$$x_{\min} \le x(t) \le x_{\max}, \quad (6d)$$
$$l_{\mathrm{on}}(t) - l_{\mathrm{on}}(t-1) \le l_{\mathrm{on}}(\omega_{\mathrm{up}}), \text{(Off/On switch)} \quad (6e)$$
$$l_{\mathrm{on}}(t-1) - l_{\mathrm{on}}(t) \le 1 - l_{\mathrm{on}}(\omega_{\mathrm{down}}), \text{(On/Off switch)} \quad (6f)$$

$\forall j \in \{\mathrm{TABS},\mathrm{FC}\}$ and $l \in \{\mathrm{CH},\mathrm{HP}\}$, where $u_{\min}$ and $u_{\max}$, respectively $x_{\min}$ and $x_{\max}$, denote the the capacity limits of the manipulated variables and the physical bounds of the stratified storage. The constraints on latency times

with minimum up and down times are given in Eq. (6e) and Eq. (6f) with $\omega_{\text{up}} = t, t+1, ..., \min(t+Np, t+T_l^{\text{up}}-1)$ if we consider the minimum up time and $\omega_{\text{down}} = t, t+1, ..., \min(t+Np, t+T_l^{\text{down}}-1)$ in the other case. Eq. (6b) denotes that at each time only one hybrid mode can be active. Since the objective given in Eq. (5) is quadratic the resulting optimisation problem is a mixed-integer quadratic problem (MIQP) to be solved each time step $t = 0, \cdots, Np-1$. For this work a branch and bound algorithm is applied, (Mayer et al., 2016), which relaxes the original problem by replacing integrality constraints, (6a), i.e. integer and particularly binary variables are allowed to span over the whole continuous interval, aiming at solving many QPs and searching for the best solution.

5 Simulation Results

In this Section the simulation setup is given. Furthermore, simulation results showing the performance of the MPC-C are discussed and the comparison results to a simuated rule-based controller (RBC) are given.

5.1 Simulation Setup

The simulation for the demonstration building described in Section 2.1 is based on a snapshot of historic data of the implemented automation system for the disturbances ambient temperature and ϑ_{G}. The simulation is shown from the beginning of September until the end of November 2014. The comparison between MPCC and RBC is given for all seasons 2014. The prediction horizon Np for all predictive controllers is 12 hours and the sampling time is one hour. The return water temperature for the cooling FC circuit, T_{FCr}^{-}, is $20°C$, whereas for the heating circuit the return water temperature, T_{FCr}^{+}, is $24°C$. The two stratified storage tanks have a volume of almost $16m^3$ each.

5.2 Demonstration of MPCC Performance

The simulation study shows the performance of the proposed modular MPC concept for the demonstration building. For this work, the power demand trajectories are given by the OL predictive controller introduced in (Killian et al., 2015) for both TABS and FC supply. Since there is only one piping system for TABS, the predicted power is split into a negative cooling, $\dot{Q}_{\text{TABS}}^{\text{ref,-}} = \min(\dot{Q}_{\text{TABS}}^{\text{ref}}, 0)$, and a positive heating trajectory, $\dot{Q}_{\text{TABS}}^{\text{ref,+}} = \max(\dot{Q}_{\text{TABS}}^{\text{ref}}, 0)$. For the FC system the reference trajectories are already split by the OL controller. The simulation for all MPCs, the LMPCs as well as the MI-MPCs, is run simultaneously. However, for the FC cooling supply two circuits are implemented but only one can be active at a time (either chiller or free cooling). Therefore, an additional rule is applied which prefers the usage of the renewable energy source respectively of the corresponding LMPC if the ambient temperature does not exceed $18°C$.

Figure 3 shows the ambient temperature profile for the simulation period from September to November 2014. One can see a drop of the mean temperature after hour 1200 by about $7°C$. The green line depicts the technical limit for the usage of the free cooling circuit.

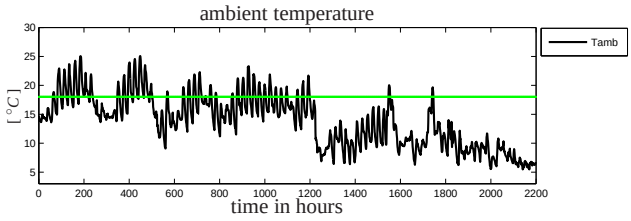


Figure 3. Ambient temperature profile starting with 1st Sept. 2014.

Figure 4 shows the simulation results for the TABS circuits, where according to the ambient temperature profile in the first half of the simulation period cooling is requested by the OL, represented by $\dot{Q}_{\text{TABS}}^{\text{ref,-}}$. The temperature spread of ϑ_{G} is sufficient to provide the required cooling energy only from the geothermal source, the manipulated variable $\dot{m}_{\text{G}}$ for the supply of $\dot{Q}_{\text{TABS}}$ is given in the second subplot. The heat pump is inactive over large parts of this first simulation period which can be seen at the red line in the last subplot. From hour 1200 onwards heating is requested by the OL, $\dot{Q}_{\text{TABS}}^{\text{ref,+}}$, and supplied by the heat pump circuit with perfect fit. The temperature of the water supply from the heat pump as well as the pumps of the heat pump circuit are presented in the second subplot.

Figure 5 shows the simulation results for the FC circuits. The references for cooling, $\dot{Q}_{\text{FC}}^{\text{ref,-}}$, and heating, $\dot{Q}_{\text{FC}}^{\text{ref,+}}$, are computed by the OL controller. The first subplot shows the performance of the MI-MPC and LMPCs for the FC circuits. In the second and third subplot the manipulated variables are presented. The red lines denote the contributions to FC heating supply of the district heat circuit, whereas the blue and grey lines show the temperature and mass flows for the chiller circuit. The green line in subplot three depicts the mass flow of the pump from the cooling tower of the free cooling circuit. The last subplot indicates which cooling circuit is active. Note, that free cooling is always active if the technical condition $T_{\text{amb}} \leq 18°C$ is fulfilled. Since the energy demand for the cooling FC circuit is low, the chiller is not regularly active, even if free cooling cannot be activated. Due to the specific hydraulic architecture of the demonstration building the pump supplying the cold water from the stratified storage tank has to be activated if free cooling is not active. Thus, a minimal deviation to $\dot{Q}_{\text{FC}}^{\text{ref,-}}$ has systematically be taken into account in these periods (see blue line in first subplot in Figure 5).

5.3 Comparison between MPCC and RBC

ESS are conventionally controlled by rule-based controllers (RBC). The RBC of the demonstration building is also simulated for comparison: Free cooling is also preferred but it may only be active if ambient temperature has been below $18°C$ for the past three hours. The chiller has to be switched on if the low storage temperature is higher than $12°C$ and switched off if it is lower than $7°C$.

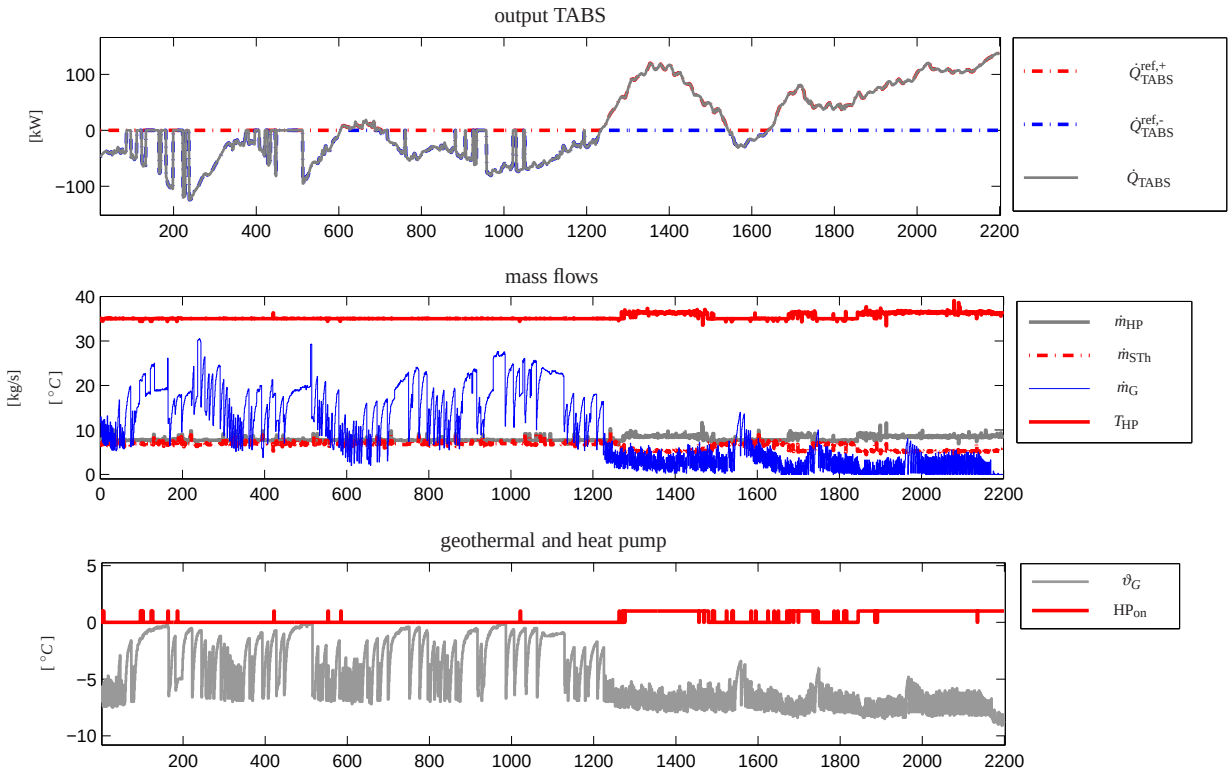


Figure 4. Simulation results for the TABS circuits starting with 1st Sept. 2014.

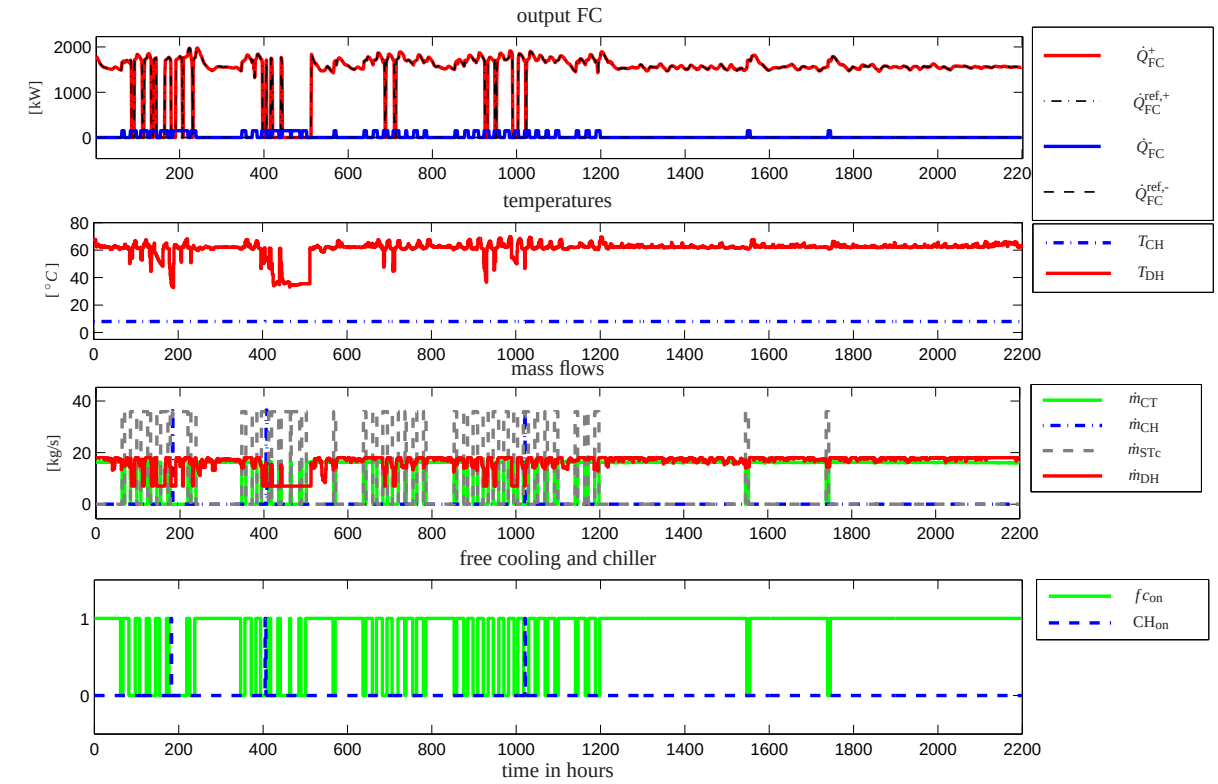


Figure 5. Simulation results for the FC circuits starting with 1st Sept. 2014.

Compared to this RBC the proposed MPCC achieves an increase of free cooling hours of almost 5% in winter (January-March and December 2014) up to over 30% in summer (June-August 2014). About 30% to 40% more cooling power $\dot{Q}^{-}$ is supplied via free cooling using the MPCC. The number of transitions of the chiller from state off to on can be reduced by around 15% in summer up to 60% in transition period (April-May and September-November 2014), which is equivalent to a significant reduction for maintenance cost for the aggregate. Table 2 shows the precise figures of the MPCC in percentage of the RBC. Simulation results show the best results for the

Table 2. Performance of the MPCC compared to the RBC in percentage according to the seasons.

season	fc hours	$\dot{Q}^{-}$ via fc	chiller state trans.
winter	104.7	130.5	0.0
transition	108.4	134.9	39.1
summer	132.6	141.2	84.8

MPCC cooling supply during summer, where the predictive character of the control concept is most effective compared to the conventional controller.

Relaxing the upper limit of $T_{\text{amb}} \leq 18^{\circ}C$ for the extended usage of free cooling leads to an average increase of mass flow of the cooling tower $\dot{m}_{\text{CT}}$ at lower variance. Table 3 lists the simulation results for the upper limit varying from $18^{\circ}C$ to $22^{\circ}C$ for autumn 2014. The number of free cooling hours is given in percentage of the simulation hours.

Table 3. Effect of a relaxation of the upper limit T_{amb} on $\dot{m}_{\text{CT}}$.

upper limit	fc hours [%]	mean $\dot{m}_{\text{CT}}$	std. dev. $\dot{m}_{\text{CT}}$
$18^{\circ}C$	89.1	21.8	8.2
$20^{\circ}C$	96.2	23.4	5.6
$22^{\circ}C$	99.5	24.0	3.6

The proposed MPCC allows to pursue different strategies for heating and cooling control, depending on the parametrisation of the weighting factors of the optimisation targets. For this work emphasis was put on the maximal usage of renewable energy sources with minimum cost while assuring minimal deviation of the delivered cooling and heating power for the supply systems in order to maximise the user comfort in the OL. The simulation results show, that the proposed MPCC is able to meet the requirements of the OL predictive controller almost perfectly and to maximise the usage of renewable energy sources such as free cooling. Furthermore, maintenance cost can be reduced due to a reduction of state transitions of the switching aggregates.

6 Conclusions

This paper introduces a modular predictive control concept (MPCC) for modern energy supply systems (ESS) of large buildings with several energy sources and supply circuits. Due to its modularity the concept is flexibly applicable for industrial building's ESS control, which is shown on a demonstration building in Austria. The resulting MPCC includes linear MPCs as well as mixed-integer MPCs (MI-MPC) dedicated for the efficient control of basic circuits respectively those with switching aggregates such as chillers with active storage. The simulation study shows that the proposed MPCC is able to accurately deliver a prescribed cooling respective heating power trajectory at minimum cost and a maximal usage of renewable energy sources. The MPCC is capable to maximise the usage of free cooling and - due to the MI-MPC - to efficiently manage the switching aggregates: compared to a RBC simulation results show an increase of free cooling hours of up to 30% and a reduction of the chiller's transitions from state off to on by up to 60% aiming at a consequent decrease of maintenance cost depending on the respective season of the year. Simulation results of the MPCC are therefore very promising for the future implementation in the demonstration building.

Acknowledgment

This work was supported by the project "SMART MSR" (FFG, No. 832103) in cooperation with evon GmbH.

References

Felix Berkenkamp and Markus Gwerder. Hybrid model predictive control of stratified thermal storages in buildings. *Energy and Buildings*, 84:233–240, 2014.

Eduardo F Camacho and Carlos Bordons Alba. *Model predictive control*. Springer Science & Business Media, 2013.

Michaela Killian, Barbara Mayer, and Martin Kozek. Cooperative fuzzy model predictive control for heating and cooling of buildings energy and buildings. *Energy and Buildings*, 2016:130–140, 2015. URL doi:/10.1016/j.enbuild.2015.12.017.

Simeng Liu and Gregor P Henze. Impact of modeling accuracy on predictive optimal control of active and passive building thermal storage inventory. *ASHRAE transactions*, 110(1), 2004.

Lennart Ljung. System identification–theory for the user, 2nd edition ptr prentice hall. *Upper Saddle River, NJ*, 1999.

Yudong Ma, Francesco Borrelli, Brandon Hencey, Andrew Packard, and Scott Bortoff. Model predictive control of thermal energy storage in building cooling systems. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pages 392–397. IEEE, 2009.

Barbara Mayer, Michaela Killian, and Martin Kozek. Management of hybrid energy supply systems in buildings

using mixed-integer model predictive control. *Energy Conversion and Management*, 98:470–483, 2015. URL doi:/10.1016/j.enconman.2015.02.076.

Barbara Mayer, Michaela Killian, and Martin Kozek. A branch and bound approach for building cooling supply control with hybrid mpc. *Energy and Buildings*, 128:553–566, 2016. URL doi:/10.1016/j.enbuild.2016.07.027.

Frauke Oldewurtel, Alessandra Parisio, Colin N Jones, Dimitrios Gyalistras, Markus Gwerder, Vanessa Stauch, Beat Lehmann, and Manfred Morari. Use of model predictive control and weather forecasts for energy efficient building climate control. *Energy and Buildings*, 45:15–27, 2012.

Luis Pérez-Lombard, José Ortiz, and Christine Pout. A review on buildings energy consumption information. *Energy and buildings*, 40(3):394–398, 2008. URL doi:/10.1016/j.enbuild.2007.03.007.

Samuel Privara, Jiří Cigler, Zdeněk Váňa, Frauke Oldewurtel, Carina Sagerschnig, and Eva Žáčeková. Building modeling as a crucial part for building predictive control. *Energy and Buildings*, 56:8–22, 2013.

Jan Širokỳ, Frauke Oldewurtel, Jiří Cigler, and Samuel Prívara. Experimental analysis of model predictive control for an energy efficient building heating system. *Applied Energy*, 88 (9):3079–3087, 2011.

Study of Different Climate and Boundary Conditions on Hygro-Thermal Properties of Timber-Framed Envelope

Filip Fedorik[1] Raimo Hannila[2] Antti Haapala[3]

[1,2]Structural Engineering and Construction Technology
University of Oulu, Finland, {filip.fedorik,raimo.hannila}@oulu.fi
[3]Wood Materials Science, University of Eastern Finland
Joensuu, Finland, antti.haapala@uef.fi

Abstract

The present paper deals with a study of different climate effects and defining boundary conditions on mould growth risk inside building envelope. The case structure represents a common envelope of timber-framed single-family house. Weather conditions from Utsjoki, Oulu and Joensuu are considered in the analysis representing climate gradients wet and dry, coastal and inland conditions during a period of 6 years.

Mould growth initiation and progression require a sufficiently high humidity at suitable temperature range. Coastal regions characteristically have humid and warm climate that causes higher risk for mould growth than the more dry inland locations. The most unfavorable conditions for mould growth were seen in the coldest and the northernmost location. Hygro-thermal simulation also presented significant differences in key interior boundary conditions that, considering standard approach, may be interpreted as potential structural health issues.

Keywords: hygro-thermal simulation, mould growth risk assessment, climate effect, structural health

1 Introduction

The recent studies in sustainability, energy use and health of buildings show a trend towards low-energy housing solutions and material development that supports energy conservation. The industry tries to create a space for the creativity of designers in designs while ensuring low total energy use, energy harvesting options and low level of heating energy losses. Although indoor designs and visible surfaces are often highlighted in house designs, characteristics of climate conditions inside the structural elements should not be forgotten, even if they are hidden from the residents.

Current structures tend to be highly insulated causing a significant gradient in the hygro-thermal conditions inside the multi-layer building components (Fedorik *et al*, 2015). A certain combination of temperature and humidity exposure leads to favorable conditions for mould growth (Hukka et al, 1999; Viitanen *et al*, 2007; Viitanen *et al*, 2010) and the existence of mould spores may cause allergic reactions (Mundarri and Fisk, 2007). The presence of mould and fungi significantly influence material properties and in long run may cause structural deterioration and costly rebuilds. This is a problem especially in cases when moisture is enclosed to envelope in building stages and the insulated wall is unable to balance it with outdoor or indoor conditions.

The presence and impact of e.g. moulds in workplaces, schools and houses has become a widely discussed issue during the recent decades. The mould issue in this context was initially found in historic buildings, where indoor ventilation was not provided or it was very inefficient (Pirinen 2011). While improved ventilation and air-exchange may indeed significantly affect the indoor conditions and protect interior wall-surfaces from the mould growth, the multi-layered envelope and/or foundation and roofs may still be under a risk of mould growth. Insufficient knowledge of designers and house owners further increase the risk of unsuitable design and maintenance issues.

There are three ways to analyses house designs for their hygro-thermal properties prior to the build-up stage: lab-tests, on-site measurements and numerical simulation. Although numerical simulation is not as accurate as the other methods, its advantage consists especially of timesaving and low expenses. If suitable verification is done, simulation is helpful in predicting the risk factors against moulds (Woloszyn and Rode, 2008).

2 Objectives

The aim of the this study consists of finding the differences in hygro-thermal conditions inside an envelope element and defining key differences in mould index (mould growth probability indicator) depending on the defined outdoor and indoor climate conditions. Data defining outdoor boundary conditions were obtained from official weather stations located in Joensuu, Utsjoki and Oulu for a 6-year period (2009-2015, see Figure 1). The indoor boundary conditions were taken as multiplication of 1-year data measured in a corresponding family house and, for comparison, data

according EN ISO 13788:2012 dependent on actual exterior conditions.

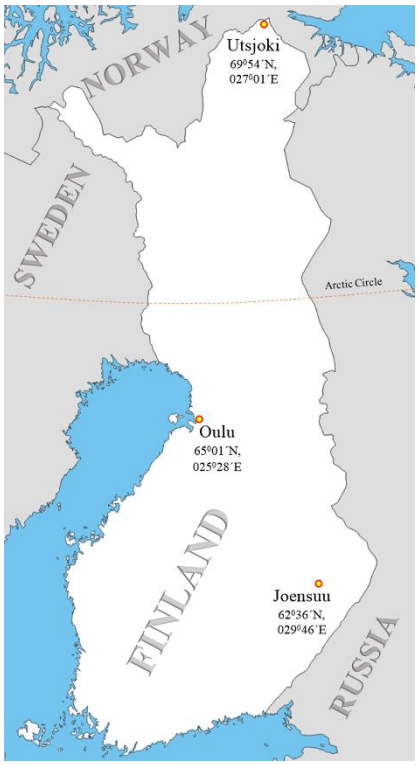


Figure 1. Localization of analyzed weather stations.

3 Applied Methodology

Temperature and humidity during the exposure time are the major factors driving the mould growth. Mould growth risk simulation is based on the model developed by the Tampere University of Technology and Technical Research Centre of Finland (VTT). RH_{crit} determines the limits between favorable and unfavorable conditions, in temperature range from 0 to 50 °C. The minimum humidity for a mould spore to start growing depends on material sensitivity, being here 80 or 85% (Ojanen *et al*, 2011). In the case the hygro-thermal conditions are favorable, the mould growth risk increases. Otherwise, if the conditions are unfavorable for mould growth initiation, the risk decreases depending on the exposed time for unfavorable conditions.

The risk of mould growth is analyzed using a mould index value (MI, M_{index}), which is expressed by an empirical model that is based on the variables of temperature, relative humidity and exposure time, as shown in our recent study (Fedorik *et al*, 2015). A border between favorable and unfavorable conditions for mould-growth initiation is defined by critical relative humidity (RH_{crit}).

The RH_{crit} varies with regard to the sensitivity of material (Ojanen *et al*, 2011), where examples and classification of materials applied in the research is specified in Table 1. The equation is valid for material sensitivity classes 1 and 2. In the case of classes 3 and 4, the minimal RHcrit must be at least 85%, which is the degree of humidity required for initiation of mould growth on a surface of this type of material.

The numerical hygro-thermal simulation was performed by Wufi®2D (Sedlbauer *et al*, 2001; Künzel, 1995), where two-dimensional model of the presented wall-envelope was created. Wufi applies two-dimensional heat and moisture transfer by the governing transport equations, while drivers for the heat and moisture transfer are vapor pressure and moisture content.

4 Analyzed Structure

The studied structure represents a common Finnish single-family timber-framed wall element. The geometry consists in gypsum board layer installed on the interior surface, 50 mm continuous mineral wool, 200 mm mineral wool filling the timber-frame, wind-proof board, 32 mm air-gap and 23 mm exterior cladding made of painted wood panels (Figure 2).

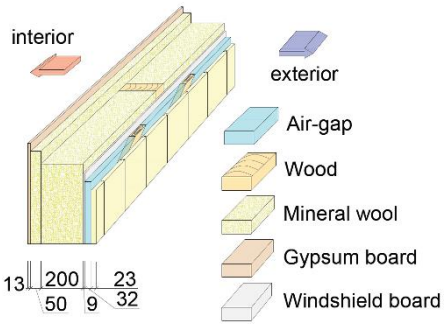


Figure 2. Schematic of the studied wall element structure.

The numerical model represents a horizontal cut-off plane of the structure (Figure 3). The hygro-thermal data was collected from 9 points (in red) considered as the most important, where humidity can creep into the structure. The geometry of the numerical model and localization of the analyzed points are given in the Figure 3. Material properties of each element are considered as they are defined in Wufi®2D material library.

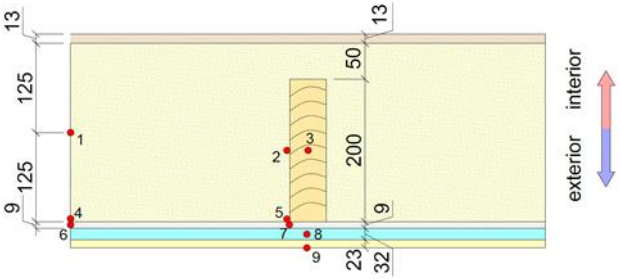


Figure 3. Analyzed points 1-9 within the wall structure.

5 Boundary Conditions

The presented study applies exterior weather conditions from three different cities across Finland. The conditions from a typical coastal, inland and northern city were applied (Figure 1). The northernmost city Utsjoki in Finland is characteristic by dry and cold weather, Oulu for its humid conditions and Joensuu represents conditions for common Finnish inland city where the climate is more warm and dry. The climate data was recorded in these cities by Finnish Meteorological Institute's weather stations for 6 years from 1.4.2009 to 31.3.2015. The weather conditions in 2010 for the cities are shown in the Figure 4 (Temperature) and Figure 5 (relative humidity. RH). The time step applied in the presented study is 1 hour.

The house interior data was considered from two perspectives. At first, as defined according to EN ISO 13788:2012 considering humidity class 3 and constant indoor temperature (at 20 °C). Secondly, actual indoor boundary conditions were measured during one year (2010) in a family house located in Oulu/Finland. These were considered to remain the same every year for the whole 6 years period.

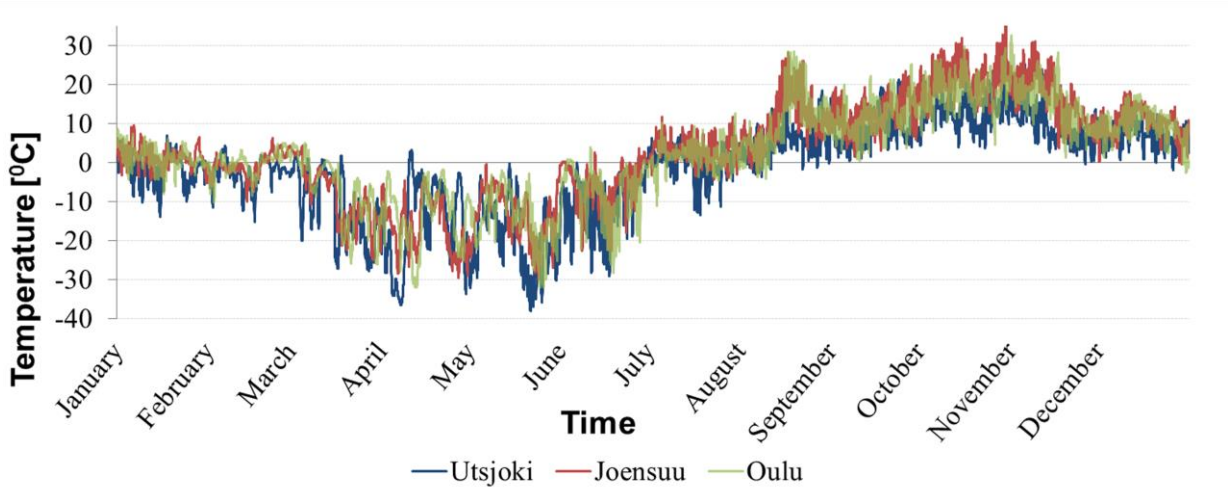


Figure 4. Development of annual temperature in 2010.

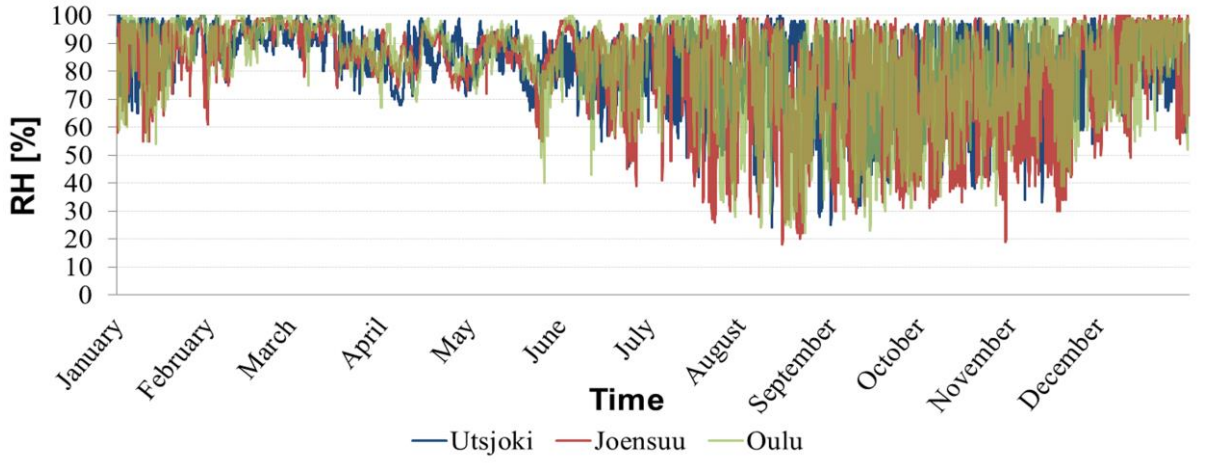


Figure 5. Development of annual humidity in 2010.

6 Results and Discussion

The presented study analyses the effect of different boundary conditions on mould growth risk inside an envelope of a common Finnish family house. The achieved results do not represent any major risk for the structure from the mould index point of view, although the differences between different cities vary significantly.

The evaluated mould index achieves quite low values, representing no or very subtle mould existence according to (Ojanen *et al*, 2011). On the other hand, significant differences were achieved in number of favorable conditions between the analyzed cities. The favorable conditions behind a possibility of mould growth initiation should be considered in house design.

Graphical illustration of favorable and unfavorable conditions for mould growth obtained at point 5, located between insulation and wind-proof board near the exterior wall-surface, are shown in Figures 6 to 8. The dots in the figures represent conditions of temperature and humidity achieved at each time-step of the solution, where the red dots represent conditions favorable for mould growth and blue unfavorable. It must be noted that the red dots do not mean initiation of mould, but they indicate the favorable conditions where growth can take place. Important part for mould starting to grow is the exposure time.

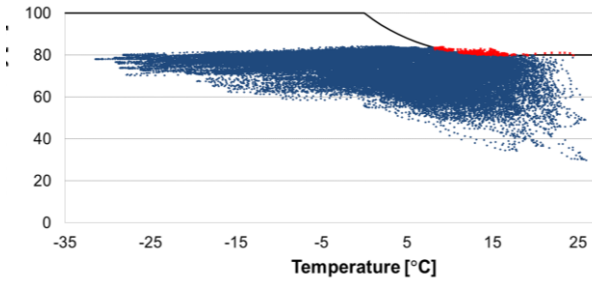


Figure 6. Favorable conditions for mould growth at exterior corner of timber-stud (point 5 in Figure 3) in Utsjoki.

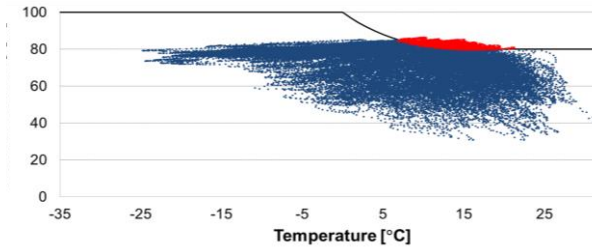


Figure 7. Favorable conditions for mould growth at exterior corner of timber-stud (point 5 in Figure 3) in Joensuu.

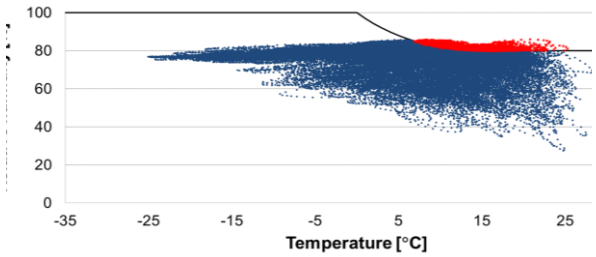


Figure 8. Favorable conditions for mould growth at exterior corner of timber-stud (point 5 in Figure 3) in Oulu.

It can be seen that the weather conditions in Utsjoki cause only briefly the conditions for favorable environment for mould growth compared to conditions in Joensuu and especially in Oulu. The evaluated maximum mould index achieved and percentage period when the structure is exposed to favorable conditions for mould growth at each of the analyzed points considering measured interior boundary conditions are summarized in the Table 1, where points 1 to 3 are not included as far as the mould index was always 0.

Table 1. M_{index} and Favourable Conditions Considering the Measured Interior Conditions.

Analysis point	Joensuu		Oulu		Utsjoki	
	$max\ M_{index}$	F_{cond} [%]	$max\ M_{index}$	F_{cond} [%]	$max\ M_{index}$	F_{cond} [%]
4	0.01	6.16	0.01	6.70	0.01	1.93
5	0.14	4.35	0.18	5.49	0.02	0.58
6	0.01	6.76	0.02	7.44	0.01	2.23
7	0.16	4.77	0.17	6.06	0.02	0.91
8	0.40	15.34	0.67	18.24	0.14	7.14
9	5.96	26.42	5.99	29.99	5.59	16.67

In the case of applying interior boundary conditions according to EN ISO 13788:2012, the favorable environment for mould growth and evaluated mould growth risk values are slightly bigger. It means that according to presented study, the applied measured interior conditions ensure healthier conditions for the structure although in both cases the mould index represents no mould growth during the analyzed period at all inner points. The summary of the favorable conditions and maximal value of mould growth risk achieved during the analyzed period is given in the Table 2.

Table 2. M_{index} and Favourable Conditions Considering Standardized Interior Conditions.

Analysis point	Joensuu		Oulu		Utsjoki	
	$max\ M_{index}$	F_{cond} [%]	$max\ M_{index}$	F_{cond} [%]	$max\ M_{index}$	F_{cond} [%]
4	0.02	11.07	0.03	13.43	0.01	3.56
5	0.27	7.01	0.39	9.79	0.03	1.58
6	0.02	10.92	0.03	13.17	0.01	3.59
7	0.22	6.12	0.29	8.26	0.03	1.40
8	0.45	15.75	0.75	18.76	0.14	7.38
9	5.96	26.42	5.99	29.99	5.59	16.67

The highest mould index value was found on the exterior surface of the envelope. A development of mould index at point 8, located in the exterior air-gap between wind-proof board and exterior cladding are illustrated in the Figure 9.

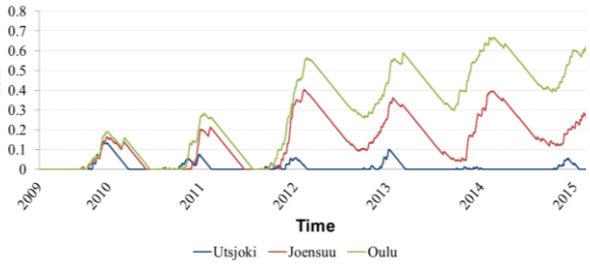


Figure 9. M_{index} in air-gap between the gypsum-board and exterior cladding (point 8 in Figure 3).

The highest mould index value was achieved on the exterior surface of the envelope. Differences in mould index development on the exterior surface of the envelope for the analyzed cities are figured in the Figure 10. At this point, notable differences in climatic variation can be seen. The humid climate in Oulu causes steeper developing of the mould index than Joensuu and Utsjoki climate. The lowest mould index is achieved in Utsjoki. It seems that for this 6-year period of analysis the houses built in all three locations are able to dry once wetted and the periods of high humidity are insufficient to cause significant issues. The ability of wall structure to dry once wetted is therefore critical to prevent significant problems and mould growth, as shown in our previous study (Fedorik *et al*, 2015).

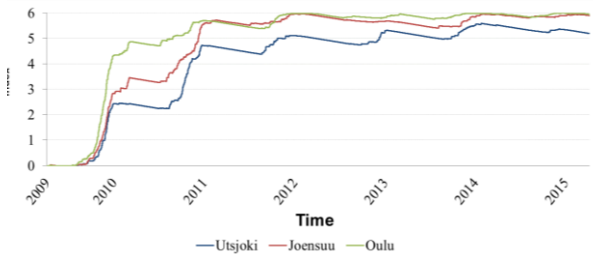


Figure 10. M_{index} on the exterior surface of envelope (point 9 in Figure 3).

7 Conclusions

The presented study presents analysis of climate effect on mould growth risk inside an envelope of timber-framed family house. As expected, the worst conditions for house health are achieved in coastal city of Oulu that has characteristically humid climate. The best results were found on climate located in the northernmost city of Finland Utsjoki, where low humidity and temperature cause unfavorable conditions for mould growth initiation.

Significant difference in numerical simulation of hygro-thermal conditions of buildings is made by defining the relevant boundary conditions. In addition, applying measured or standardized interior boundary conditions lead to different results, although none of them represents a danger of serious mould growth in the analyzed case. On the other hand, if more sensitive materials would be used, mould growth in a structure in humid environment may occur. Using one set of conditions to analyze a standard structure is likely to lead into misjudgment of the structural health if the actual location of the house is significantly different in its climate conditions.

References

F. Fedorik, M. Malaska, R. Hannila, and A. Haapala. Improving the thermal performance of concrete-sandwich envelopes in relation to the moisture behaviour of building structures in boreal conditions. *Energy and Buildings*, 107: 226–233, 2015.

A. Hukka and H. A. Viitanen. A mathematical model of mould growth on wooden material, *Wood Science and Technology* 33(6): 475–485, 1999.

H. M. Künzel. *Simultaneous heat and moisture transport in building components*. Fraunhofer Institute of Building Physics, Germany, 1995.

D. Mudarri and W. J. Fisk. Public health and economic impact of dampness and mold, *Indoor Air Journal*, 17: 226-235, 2007.

T. Ojanen, R. Peuhkuri, H. Viitanen, K. Lähdesmäki, J. Vinha, and K. Salminen. Classification of material sensitivity – New approach for mould growth modelling. *9th Nordic Symposium on Building Physics*, 29 May – 2 June 2011, Tampere, Finland.

J. Pirinen. Building inspections in Finland – Fighting against the moulds, moisture and mould programme. *9th Nordic Symposium on Building Physics*, 29 May – 2 June 2011, Tampere, Finland.

K. Sedlbauer, M. Krus, W. Zilling, and H. M. Künzel. Mold growth prediction by computational simulation. *ASHRAE Conference*, IAQ, San Francisco, CA, 4-7 November 2001.

H. A. Viitanen and T. Ojanen. *Improved model to predict mold growth in building materials*. Report based on the VTT projects "Building Biology" and "Integrated Prevention of Moisture and Mould Problems", 2007, Finland.

H. Viitanen, J. Vinha, K. Salminen, T. Ojanen, R. Peuhkuri, L. Paajanen, and K. Lähdesmäki. Moisture and bio-deterioration risk of building materials and structures. *Journal of Building Physics*, 33(3): 201–224, 2010.

M. Woloszyn and C. Rode. Tools for performance simulation of heat, air and moisture conditions of whole buildings. *Building Simulation*, 1(1): 5–24, 2008.

Evaluation of Structural Costs in Building - Simulation of the Impact of the Height and Column Arrangement

Javier Ferreiro-Cabello [1] Esteban Fraile-García [1] Eduardo Martínez de Pisón-Ascacíbar [1]
Emilio Jiménez-Macías [2]

[1] Department of Mechanical Engineering, University of La Rioja, Spain, {javier.ferreiro,esteban.fraile, eduardo.mtnezdepison}@unirioja.es

[2] Department of Electrical Engineering, University of La Rioja, Spain, emilio.jimenez@unirioja.es

Abstract

Modeling is a useful tool for decision making in the project phases. In the case of reinforced concrete structures, we must be able to locate representative parameters in order to optimize costs. This paper assesses the impact of the column arrangement and building height. The variation of the costs for the foundation and two floor interaxis are discussed. The results are assessed by the ratio of cost per square meter executed. The optimization of the geometry of the building is determined by the interaxis distances and the selected structural thickness. In the case studied the arrangement of the pillars in a 6x6 meters grid using 4 heights offers the best economic results.

Keywords: reinforced concrete, costs, columns arrangement, structures, modeling

1 Introduction

The decisions made in the early stages of a project have a significant impact on its future development. The economic costs can and should be dimensioned so that the investor has the smallest possible uncertainties. Despite this construction projects have a long deployment time that introduces those unwanted uncertainties. This work focuses on structural solutions through the use of reinforced concrete (Amir, 2013; CTE, 2006; CYPECAD, 2015).

The intensive application of reinforced concrete has been produced largely by the advancement and study of the behavior of new materials and the development of new technologies. This is the origin of structural engineering that deals with the conception, design and construction of structures emerged. We want to emphasize that in the same way that society evolves, technology, materials and available tools do (De Albuquerque *et al*, 2012; Delijani *et al*, 2015; EHE-08 2008).

The structural solution costs represent a significant percentage within any project. With an eye to the future by implementing algorithms it is necessary to know the impact of different variables that affect the final cost of a given solution (Fernández-Ceniceros *et al*, 2010; Kaveh *et al*, 2011; Koksal *et al*, 2013).

As representative geometric variables in the definition of one building they have been considered the number of pillars or columns and the height of the building. For the structural solution of the floors a bidirectional forged or slab structure recoverable coffer with a constant structural depth has implemented, but modifying the interaxis (Moretti, 2014; Poluraju *et al*, 2012; Porwal and Hewage, 2012).

This range of solutions is made by using three elements mainly: Concrete, steel and formwork elements.

2 Proposed Methodology

The proposed methodology focuses on assessing the economic impact incurred in the process of building a reinforced concrete structure assuming that the reticular forged recoverable coffer is selected for the horizontal structure. This choice is not accidental because, it presents some remarkable features:

- Materials incorporated to the structure are permanently only two, in this case steel and concrete. In all cases it is used concrete HA-25/P/20/IIa and steel B-500S.
- In this case the difference in the performance of each alternative is faithfully reflected in the variation of the quantities consumed of steel and concrete.
- The use of recoverable coffer.

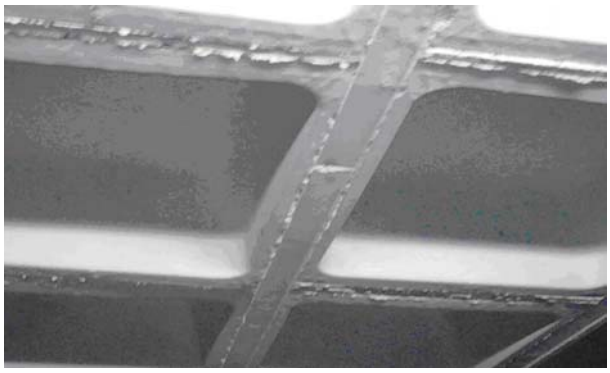


Figure 1. Final result after removal of the provisional formwork.

To execute these structures, once consolidated the pillars, a framework is used (Figure 1). Once the structure consolidates these elements are retrieved and used in subsequent structures. For the assessment of costs it is important to establish the number of uses. In this paper this variable is 50.

Obviously the cost of a structure will be lower when the material consumption are optimized (reducing the amounts of concrete and steel) and the use of formwork materials and labor.

The proposal is to make the modeling of a square building of dimensions 24x24 meters using different arrangements of columns and number of plants (Figure 2). For the arrangement of pillars three values of the grid have been selected: a situation of short lights of 4x4 meters, another common situation in building alternative of 6x6 meters, and 8x8 meters with overhead lights. The modeled cases include 4 floors, 6 floors, 8 floors and 10 floors. The buildings have a height on the ground floor of 4 meters and the rest of floor slabs with heights of 3 meters, devoting the last forged to a flat roof.

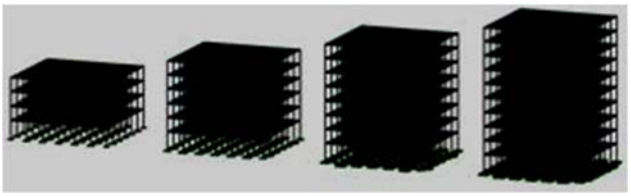


Figure 2. Modelling buildings.

All the floors have been solved using a structural depth of 30 centimeters (25 of box more 5 cm of compression layer). The distances of the models are varied using two discrete solutions of 60 and 80 centimeters (Figure 3), the width of nerve and the coating remained constant. Thus the own weights of the two alternatives of the implemented floor slab are 4.70 kN/m^2 and 4.03 kN/m^2 respectively.

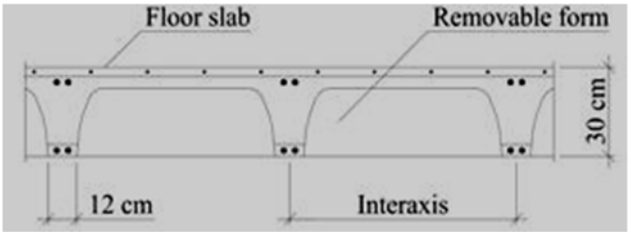


Figure 3. Alternative sections of implemented floors.

Regarding the considered loads, facade loads have been introduced as uniform loads on the perimeters of the floors with a value of 7 kN/m, and on deck this value is reduced to 3 kN/m. For surface loads on the floors 2kN/m^2 has been considered for permanent loads and 2kN/m^2 for overhead use. In the cover these values have changed, 3 kN/m^2 for permanent loads and 1kN/m^2 for overload use. Wind loads were implemented considering the Spanish legislation and snow loads are included in overload considered indoor use. For dimensioning the foundation, it has been considered an average benefit in the soil bearing capacity, on the maximum permissible stress, 0.2 N/mm^2.

Table 1. Unit Cost Items Considered.

Description	**Cost (€)**
m² System formwork foundation plinth.	19.94
m² lean concrete layer (thickness 0.1 m).	10.22
m^3 foundation of reinforced concrete, concrete made with HA-25/P/20/IIa manufactured in plant, and discharge from truck.	104.70
Reinforcing steel kg UNE-EN 10080 B 500 S, developed in industrial workshop. Including transportation and placement work.	1.00
m^3 of concrete for pillars made of concrete HA-25/P/20/IIa manufactured in central and poured with cupolas, assembly and disassembly of reusable formwork system metal sheets.	349.65
Reticular m², total depth 30=25+5 cm, made with concrete HA-25/P/20/IIa manufactured in central; discharge with pump on continuous formwork system; nerves "in situ" 12 cm, welded wire in compression layer. No impact of pillars.	37.60
Unites of recoverable formwork PVC, 76x80x25 cm for 50 uses, including special pieces.	2.29
Unites of recoverable formwork PVC, 56x60x25 cm for 50 uses, including special pieces.	1.75
m^3 of concrete for slabs manufactured in Central HA-25/P/20/IIa.	76.88

The definition of the structure will be made following the Spanish legislation and using a structural calculation software tool named CYPECAD. Performing calculations provides data on the consumption of materials, which in the selected type represent significant values used in the comparison. By using a database of construction, the prices of each of the studied alternatives are obtained.

Table 2. Items Considered for Each Block Of The Structure.

	items	**units**
foundation	Cleaning concrete HL-15 / P / 20	m^3
	Reinforcing steel B 500 S	kg
	Concrete HA-25 / P / 20 / IIa	m^3
	Shuttering fundation	m^2
columns	Column formwork	m^2
	Reinforcing steel B 500 S	kg
	Concrete HA-25 / P / 20 / IIa	m^3
floor	Formwork wrought	m^2
	Reinforcing steel B 500 S	kg
	Concrete HA-25 / P / 20 / IIa	m^3
	boxes	units

The cost of the structure is divided into three sections: foundation, pillars or columns and floor. This scheme follows the construction process, and the prices for the various items are presented in Table 1.

These prices combined with the results of consumption of each alternative allow us to obtain the costs of the proposed solutions. Table 2 lists the items that are incorporated in each block with the units used.

Table 3. Consumption Obtained for Each Solution

			consumption Foundation				*consumption pillars*			*Forged consumption*		
I (cm)	**R (mxm)**	**H (nº)**	**HL (m^3)**	**Fe (kg)**	**HA (m^3)**	**E (m^2)**	**E (m^2)**	**Fe (kg)**	**HA (m^3)**	**Fe (kg)**	**HA (m^3)**	**C (Ud)**
60	4X4	4	15	2196	55	125	694	4238	51	16213	537	4224
		6	23	3761	116	210	1011	7040	75	24841	805	6336
		8	32	6347	238	368	1341	11573	100	34570	1073	8448
		10	41	1068	366	499	1674	22563	126	46543	1341	10560
	6x6	4	14	2660	96	151	360	3507	27	23555	497	5184
		6	22	5307	201	260	539	8792	42	37219	746	7776
		8	30	9095	307	344	736	16570	60	53793	994	10368
		10	38	11921	397	396	965	24209	85	75531	1242	12960
	8x8	4	14	3659	143	182	250	5538	21	40697	492	5272
		6	22	6752	232	236	383	13554	34	63716	738	7908
		8	31	11011	356	307	541	20154	52	91923	984	10544
		10	38	14444	463	356	720	27793	76	125531	1230	13180
80	4X4	4	16	2279	62	137	694	4238	51	16544	576	1512
		6	24	4149	127	226	1011	7350	75	25603	864	2268
		8	33	6389	264	402	1347	12259	101	36131	1151	3024
		10	42	11294	372	503	1678	23405	127	49036	1439	3780
	6x6	4	14	2385	89	145	360	3284	27	21524	451	2820
		6	21	5007	186	249	539	8409	42	34075	677	4230
		8	29	8314	293	335	733	15580	60	50216	902	5640
		10	37	11400	376	385	957	26114	83	70901	1128	7050
	8x8	4	14	3329	137	176	250	5322	21	39751	446	2880
		6	21	6618	217	228	385	12518	34	63391	669	4320
		8	29	10405	333	297	534	19485	51	91053	892	5760
		10	37	13885	446	347	713	26657	74	124024	1115	7200

Table 4. Costs of Different Structural Proposals

I (cm)	**R (mxm)**	**H (nº)**	**€ Foundation**	**€ Pillars**	**€ Slabs**	**€ Structure**	**€/m^2**
60	4x4	4	12.006	22.227	153.676	187.910	79,6
		6	22.359	33.253	231.036	286.648	80,9
		8	41.832	46.681	309.484	397.997	84,2
		10	53.482	66.776	390.169	510.427	86,4
	6x6	4	17.238	13.021	159.665	189.924	80,4
		6	33.829	23.554	241.381	298.765	84,3
		8	51.199	37.703	325.983	414.885	87,8
		10	65.340	53.866	415.772	534.979	90,6
	8x8	4	23.751	12.783	176.604	213.138	90,2
		6	38.042	25.428	267.555	331.025	93,4
		8	57.571	38.350	363.709	459.630	97,3
		10	73.916	54.227	465.260	593.403	100,5
80	4x4	4	13.142	22.227	153.081	188.450	79,8
		6	24.388	33.563	230.408	288.360	81,4
		8	45.407	47.703	309.187	402.296	85,2
		10	64.508	67.870	390.341	522.720	88,5
	6x6	4	16.051	12.798	151.484	180.333	76,3
		6	31.572	23.140	229.006	283.717	80,1
		8	48.681	36.521	310.108	395.310	83,7
		10	62.185	55.289	395.787	513.261	86,9
	8x8	4	22.591	12.567	169.490	204.647	86,6
		6	36.097	24.518	257.979	318.594	89,9
		8	54.148	37.247	350.499	441.894	93,5
		10	71.244	52.629	448.337	572.209	96,9

3 Objective

The main objective is to have a vision of the cost of the entire structure. The inclusion of three variables will allow to compare different alternatives and select the one that is most viable from the economic point of view.

The novelty lies on the inclusion of variations in the geometry of the building (arrangement of the pillars and building height) combined with the definition of the structural solution used (forged 60x60 and 80x80 centimeters interaxis).

To facilitate the monitoring of the cost of the structure it has been divided into three blocks with a different treatment. The foundation with four blocks for the lean concrete, steel foundation, structural concrete and formwork needed. The pillars are elements with a production process in which the cleaning concrete disappears, repeating the previous three blocks and presenting totally different yields.

4 Case study

Performing calculations by regulations currently in use in Spain determines the technically viable alternatives. From those viable solutions, material consumption are found: steel, concrete and auxiliary elements (formwork and caissons) at different values for each block of the structure; the data are reflected in Table 3. The modeling analysis allows controlling deformations and adapting the optimal arrangement of the structure, all rigorously complying with current regulations.

Known data set consumption and prices of each item can determine the cost of each alternative as reflected in Table 4.

The production cost is done add the amount of each item. In this case we have taken into account the variations in consumption of both materials and auxiliary means necessary.

The total cost of the structure has been obtained and the graphical representation of the results is divided into two blocks. Figure 4 represents the impact of each block in each studied alternative.

For this configuration of the floor inside buildings of four floors, solution employing fewer supports (8x8) has a total cost that is 13.43% higher than the alternative with the highest number of supports (4x4). In the case of solutions for buildings of 10 floors this difference becomes greater resulting in a 16.26%. If we analize the results in regard to the foundations these values are completely different in low buildings since the cost of the foundation is increased by 97.83% whereas in tall buildings the decreased number of pillars an increase of 38.21%. Reducing the number of supports makes the starting pillars decrease.

With 60x60 interaxis configuration total costs increase as the number of carriers it is reduced for all modeled buildings.

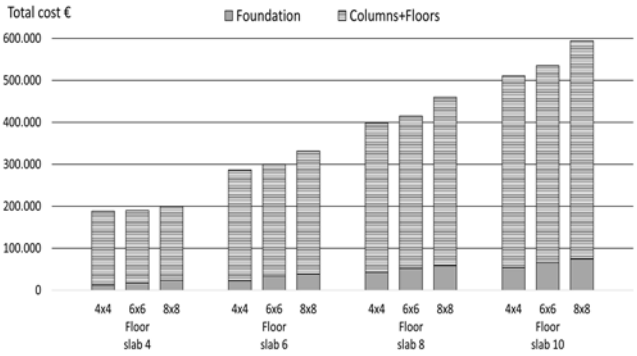


Figure 4. Total cost alternatives interaxis 60x60.

The results for structural solutions that employ larger interaxis (80x80) are shown in Figure 5.

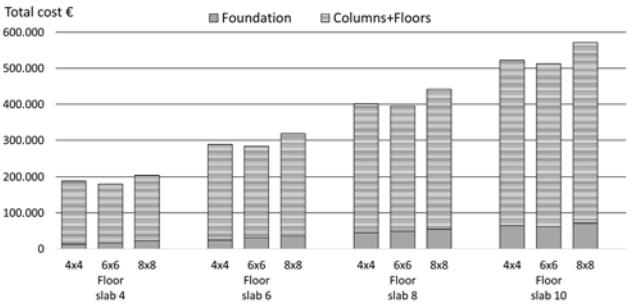


Figure 5. Total cost alternative interaxis 80x80.

For this configuration of the floor inside buildings of four floor employing fewer supports (8x8) has a total cost which is 8.59% higher than the alternative with the highest number of supports (4x4). In the case of buildings of 10 floors this difference becomes greater resulting in a 9.47%. If we analyze the results in regard to the foundations these values are completely different in low buildings since the cost of the foundation is increased by 71.90% whereas in tall buildings the decreased number of pillars drives to an increase of 10.44%. Reducing the number of supports makes the starting pillars decrease and the cost of the slabs being greater increases in recent increases.

This configuration of the floor, increasing the interaxis, allows better solutions in total cost when the 6x6 grid for the same height of the building is implemented. In the four heights minimum cost values are obtained.

Solutions with more supports (4x4) have similar costs for both interaxis being lower in 60. The differences are below 2.40%. Solutions with fewer brackets (8x8) costs are similar, being lower in 80. The resulting differences are extended below 4.01%.

While these results are interesting, we consider much more relevant to compare the ratios of the structural solution. In this case the cost per executed square meter is analyzed, and the numerical results are reflected in the

last column of Table 4, while graphical results are presented jointly for both interaxis in Figure 6.

The extreme values have a variation of 31.72% and are:

- The minimum cost 76.3 € / m2 for a building of 4 heights with me interaxis 6x6 grid 80 cm.
- The maximum cost 100.5 € / m2 for a building of 10 heights with me interaxis 8x8 grid 60 cm.

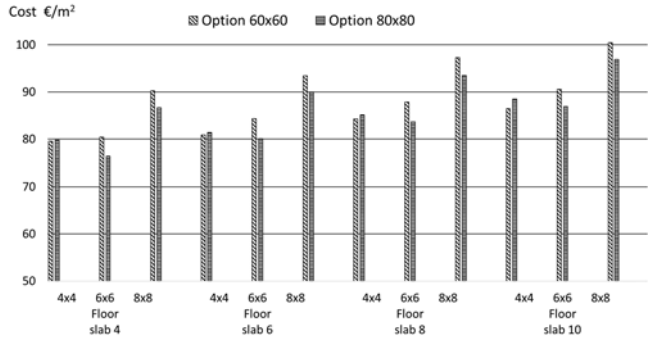


Figure 6. Ratio of cost € / m2 different alternatives.

Interaxis solutions by 60 cm range from 79.6 € / m2 for buildings of 4 heights and 4x4 grid to 100.5 € / m2 in the case of 10 heights and 8x8 grid. This is a variation of 26.26%. For proposals by interaxis 80 cm range from 76.3 € / m2 for buildings of 4 heights and 6x6 grid to 96.9 € / m2 in the case of 10 heights and 8x8 grid. This is a variation of 27.00%.

5 Conclusions

As a preliminary conclusion, it is noteworthy that variations make that the results present significant oscillations. The own reinforced concrete structural definition incorporates decisions affecting the cost produced in the design phase and implementation.

The tool implemented here is very useful when combined and incorporated the cost or impact of the land on which it is intended to build. The combination of both values allows the designer to locate a lower cost alternative.

Disregarding the impact of the land, and for a structural thickness of 30 centimeters, the most economical solutions are located in low buildings of 6x6 meters grid and interaxis distances of 80 centimeters. The worst alternative is located when employed 60 centimeters interaxis and reticles of 8x8 meters.

Structurally, the 30 centimeter thickness is oversized for reticles of 4x4 meters and it presents very high amounts of steel for reinforcement grids of 8x8 meters. This is the reason why the best solutions appear in the grid of 6x6 meters. This phenomenon is more pronounced when increasing the interaxis distances. This is one of the reasons why this is the most used structural thickness in structures in buildings in Spain.

This novelty presents a clear practical application to real cases, since the casestudy has been selected only as a way to present the proposed methodology based on the ratio of cost per square meter executed, but the methodology is widely applied to real cases. In fact, this piece of research is based on the information obtained from thousands of real cases, from a building interprise, which have also been used to validate the proposal.

Furthermore, the use of the results of this work by the designer permits the optimization of the decisions based on the conclusions obtained in the general case.

References

O. Amir. A topology optimization procedure for reinforced concrete structures. *Computers & Structures*, 114:46-58, 2013.

CTE. "*Código Técnico de la Edificación*", translated as Technical Building Code. Royal Decree 314/2006, of March 17, of the Ministry of Housing (Spain), 2006.

CYPECAD "*CYPE Ingenieros, Software para Arquitectura, Ingeniería y Construcción*", translated as CYPE Engineers, Software for Architecture, Engineering and Construction. CYPE Ingenieros S.A., 2015.

A.T. De Albuquerque, M.K. El Debs, and A.M.C. Melo. A cost optimization-based design of precast concrete floors using genetic algorithms. *Automation in Construction*, 22:348-356, 2012.

F.Delijani, M.West, and D. Svecova,. The evaluation of change in concrete strength due to fabric formwork. *Journal of Green Building*, 10(2):113-133, 2015.

EHE-08 "*Instrucción de Hormigón Estructural*" translated as Structural Concrete Instruction, Royal Decree 1247/2008, of July 18, of the Ministry of Public Works (Spain).

J. Fernández-Ceniceros, E. Martínez-De-Pisón, F.J. Martínez-De-Pisón, R. Lostado-Lorza, and L. Celorrio. "Optimización de costes en estructuras de hormigón mediante técnicas de minería de datos. Aplicación a forjados unidireccionales" translated as Cost optimization in concrete structures using data mining techniques. Application to unidirectional slabs, *XIV International Congress on Project Engineering*, Madrid (Spain), 2010.

A. Kaveh, and A. Shakouri Mahmud Abadi. Cost Optimization of Reinforced Concrete One-Way Ribbed Slabs Using Harmony Search Algorithm, *Arabian Journal for Science and Engineering*, 36(7):1179-1187, 2011.

F. Koksal, A. Ilki., and M.A. Tasdemir. Optimum Mix Design of Steel-Fibre-Reinforced Concrete Plates, *Arabian Journal for Science and Engineering*, 38(11):2971-2983, 2013.

L. Moretti. Technical and economic sustainability of concrete pavements. *Modern Applied Science*, 8(3):1-9, 2014.

P. Poluraju, K. Dorji, P. Rai, T. Wangchuk, and Tharchen. Economic design of concrete structure through judicious selection of materials at the early stage of design phase. *International Journal of Earth Sciences and Engineering*, 5(2):358-362, 2012.

A. Porwal, and K.N. Hewage. Building information modeling-based analysis to minimize waste rate of structural reinforcement. *Journal of Construction Engineering and Management*, 138(8):943-954, 2012.

Efficiency of QEs in USA Through Estimation of Precautionary Money Demand

Yoji Morita Shigeyoshi Miyagawa

Department of Economics, Kyoto Gakuen University, Japan, {morita-y,miyagawa}@kyotogakuen.ac.jp

Abstract

FRB adopted "quantitative monetary easing" three times as QE1 (2008m11,2010m06), QE2 (2010m11,2011m06) and QE3 (2012m09,2014m12). In this paper, we showed that "Reserve at the FRB" is effective to the economy through a transmission path of a stock market in QE1, effective through housing price channel in QE2 and QE3, and effective through an exchange rate channel in QE3, where impulse responses in VAR model are calculated with "reserve, stock prices, exchange rate, industrial production, and cpi_core (or housing price)" in monthly data of USA.

Furthermore, we investigated behaviors of M2 money in QEs periods. Decomposing M2 into transaction money demand and precautionary one, we estimated precautionary money demand as a function of industrial production, business condition denoted by napm and reserve at the FRB. We showed that increasing "Reserve at the FRB" is comparatively effective in QE1 rather than in QE2 and QE3 through the behavior of napm.

Keywords; QE1, QE2, QE3, nontraditional monetary policies

1 Introduction

The subprime problem in 2007 and Lehman crisis in 2008 caused serious depressions in the world economy. Many central banks set interest rates around zero, and carried out "nontraditional monetary policies" in large scales. Generally speaking, operation of short term interest rates based on for example Taylor rule is called "traditional monetary policy", while in financial crisis of these days a traditional monetary policy has no room to operate around zero interest rates, and hence, many central banks were forced to adopt nontraditional monetary policies. There are three kinds of nontraditional monetary easing in USA.

Federal Reserve Board (FRB) decreased FF rate from 2 % at Lehman crisis (2008m09) to 0-0.25 % (2008m12). Furthermore, additional easing policies were done by operations of buying long term government bond, Residential Mortgage-Backed Securities (RMBS) and agency debt. FRB called these policies as "Credit Easing" in the interval (2008m11,2010m06). We denote these monetary easing in this period as QE1. FRB carried out QE2 during (2010m11,2011m06), where long term government bonds of 600 billion dollars were purchased. QE3 was operated during (2012m09,2014m10), where FOMC decided on 2012m12 to buy MBS of 40 billion dollars and long term government bonds of 45 billion dollars per month. However, from 2014m01, FRB gradually decreased buying operations every month and stopped QE3 on 2014m10.

(Bernanke, 2009) said that the essence of QE1 is "credit easing", that is, reducing the cost of private brrowing by direct purchases of privately issued debt instead of government debt. (Gagnon et al., 2011) reported that large-scale asset purchases in QE1 have been successful in doing lower longer term private borrowing rates, which should stimulate economic activity. (Fratzscher et al., 2013) showed highly effectiveness of QE1 compared with QE2 and analysed the global spillovers of the FRB's QE. International spillover effects of US QE were investigated by (Bhattarai et al., 2015). (Wen, 2014) studied the likely impact of QE and its exit strategy with three aspects; (i)the timing of the exit, (ii)the pace of the exit and (iii)the private sector's expectations of when and how the FRB will exit. (Engen and Reifschneider, 2015) showed that by the unconventional monetary policies in USA the peak unemployment effect does not occur until early 2015, while the peak inflation effect is not anticipated until early 2016.

(Hall et al., 2012) analyzed European economy including financial crises 2007 and 2008, focusing on the stability of M3 money demand function through a generalized cointegration concept "TVC". (Fawley and Neely, 2013) described the circumstances of and motivations for the quantitative easing programs of the FRB, Bank of England, European Central Bank, and Bank of Japan during the recent financial crisis and recovery.

Efficiency of QE in Japan during (2001,2006) was shown by (Honda et al., 2007) through transmission paths of both stock market and exchange rate. They used SVAR model. (Sawada, 2014) applied Honda's method to USA case and showed that base money operated at QEs are efficient to the economic activity.

First, in this paper, following (Honda et al., 2007) and (Sawada, 2014), we construct VAR models with 5 variables (reserves, stock prices, exchange rate, industrial production, cpi) in QE1, QE2 and QE3, and investigate the efficiency from "reserves" to "industrial productions" through transmission paths of stock market, exchange rate and/or housing price in each of QEs.

Secondly, focusing our attention on money demand function of "M2", we decompose "M2" into precautionary

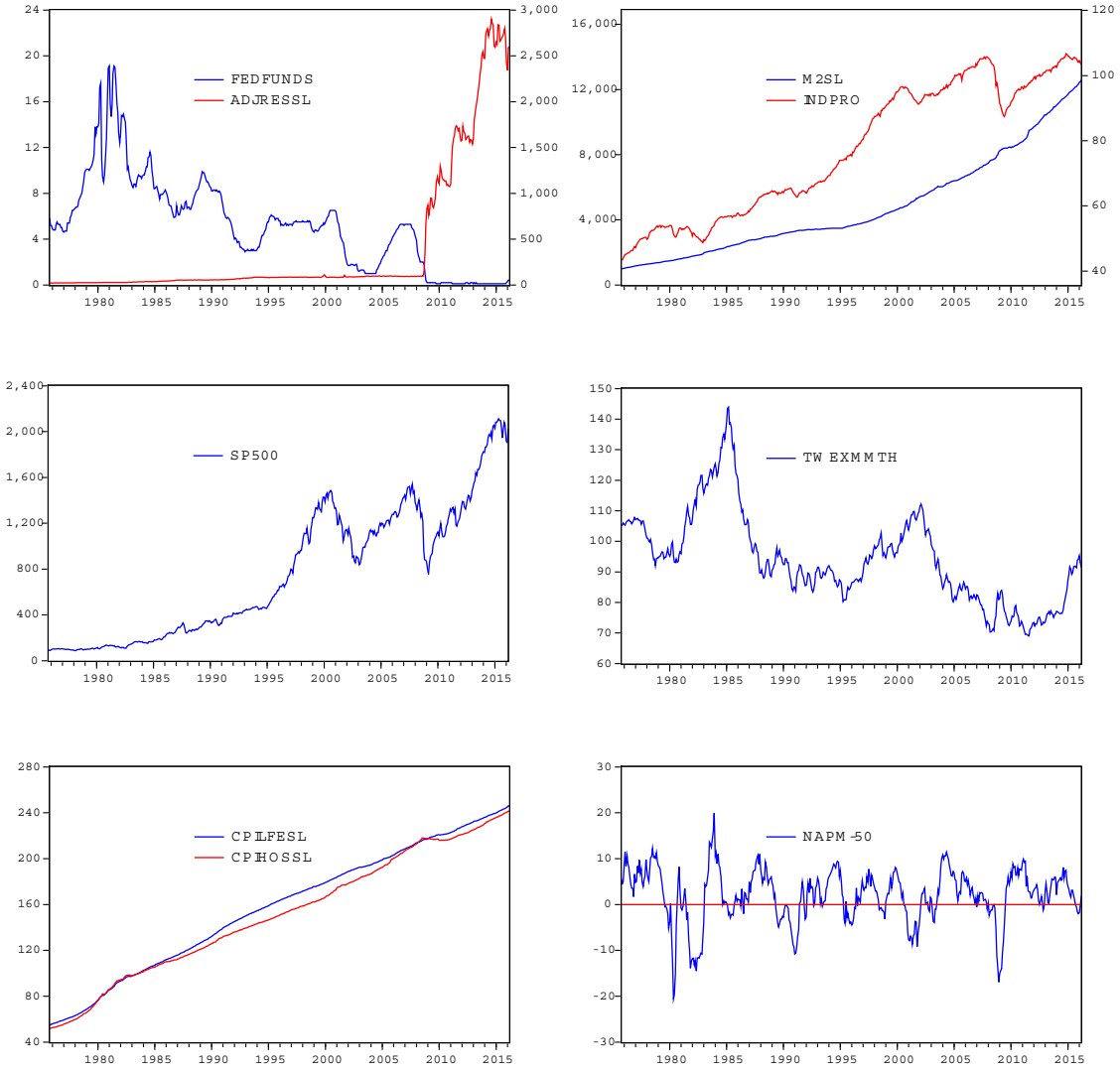


Figure 1. $fedfunds(t)$ and $adjressl(t)$, $indpro(t)$ and $m2sl(t)$, $sp500(t)$, $twexmmth(t)$, $cpilfesl(t)$ and $cpihossl(t)$, and $u(t)$ in (1975m10, 2016m03)

and transaction money demands and estimate precautionary one as a function of economic activity, reserves and business condition during (1975m10,2016m03). Thus, we can investigate behavior of M2 money in QEs. Since (Morita and Miyagawa, 2016) estimated precautionary money demand and analyzed QE monetary policy in Japan with quarterly data, we apply this method to the analysis of US data.

2 Data Properties

Monthly data through (1975m10, 2016m03) are obtained from FRED. Variables and symbolic notations are given in Table 1. See Figure1 for behavior of each variable.

Two kinds of unit root tests are carried out; DF-GLS (ERS) test with unit root as the null hypothesis and KPSS test with stationarity as the null hypothesis. The results are shown in Table 2. Every variable except $u(t)$ is shown to be nonstationary. Hereafter, we treat these variables in

Table 1. List of variables

$adjressl(t)$ = St.Louis adjusted reserves
$fedfunds(t)$ = federal funds rate
$m2sl(t)$ = M2 money stock
$indpro(t)$ = industrial production index
$cpilfesl(t)$ = consumer price index, less food & energy
$cpihossl(t)$ = consumer price index, housing
$sp500(t)$ = S&P500 index
$twexmmth(t)$ = trade weighted exchange index
$napm(t)$ = ISM manufacturing, PMI composite index
$u(t) = napm(t) - 50$
$u_-(t) = \min\{u(t), 0\}$ $\quad$ $u_+(t) = \max\{u(t), 0\}$

Table 2. Unit Root Test in $(1975m10, 2016m03)$

var.	ERS	*lag*	KPSS	*trend*
$\ln(m2sl/p)$	4.98	2	2.51∗	*const.*
$\ln(indpro)$	1.52	3	2.72∗	*const.*
$\ln(sp500/p)$	-2.02	1	0.29∗	*trend.*
$\ln(twexmmth)$	-1.32	1	1.47∗	*const.*
u	$-4.33*$	2	0.01	*const.*
$\ln(adjressl/p)$	0.92	2	1.66∗	*const.*
$\ln(p)$	0.24	9	2.61∗	*const.*
$\ln(cpihossl)$	0.61	9	2.62∗	*const.*

∗ denotes 1% significance level and $p = cpilfesl$.

levels in order to avoid cointegration analysis.

3 Macro Money Systems in QEs

Letting $x = (\ln(adjressl) - \ln(p),\ \ln(sp500) - \ln(p),\ \ln(twexmmth),\ \ln(indpro),\ \ln(p))'$, we consider VAR model of the form with the lag order i given by AIC,

$$x(t) = A_0 + A_1 x(t-1) + \cdots + A_i x(t-i) + \varepsilon(t), . \qquad (1)$$

3.1 Behavior in QE1=(2008m11,2010m06)

Setting sampling interval as QE1, impulse responses of (1) are depicted in Figure 2, where a solid line implies a calculated impulse response and two dotted lines show 95 % confidence intervals. For economy of space, we only show responses of 5 variables corresponding to three kinds of impulse shocks; ”reserves“, ”sp500“ and ”exchange rate (twexmmth)“.

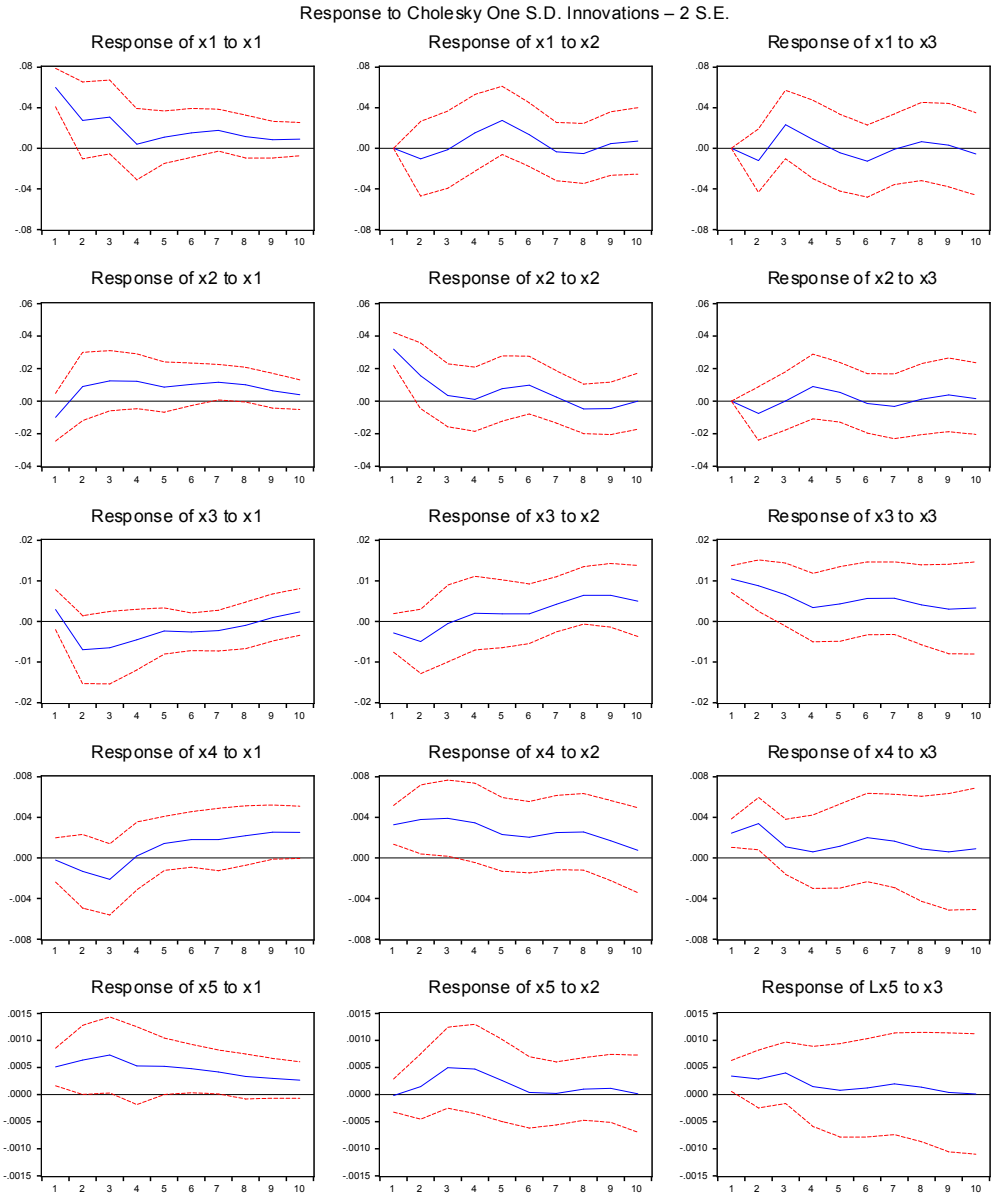


Figure 2. Impulse responses of the system with $x = (x_1, x_2, x_3, x_4, x_5)'$, where in the figure we denote $x_1 = \ln(adjressl/p)$, $x_2 = \ln(sp500/p), x_3 = \ln(twexmmth), x_4 = \ln(indpro), x_5 = \ln(p)$ in QE1=(2008m11,2011m06).

Hereafter, if necessary, we abbreviate *reserves* and *twexmmth* by *rsrvs* and *rex* respectively.

1st:	$rsrvs(\uparrow) \Longrightarrow$	$rsrvs(\uparrow)$	$stock(\cdots\uparrow)$
	$rex(\downarrow)$	$indpro(\cdots\uparrow)$	$p(\uparrow)$
2nd:	$stock(\uparrow) \Longrightarrow$	$rsrvs(\cdots)$	$stock(\uparrow)$
	$rex(\cdots)$	$indpro(\uparrow)$	$p(\cdots)$
3rd:	$rex(\uparrow) \Longrightarrow$	$rsrvs(\cdots)$	$stock(\cdots)$
	$rex(\uparrow)$	$indpro(\uparrow)$	$p(\uparrow)$

where $rex = twexmmth(\uparrow)$ implies "high appreciation of dollar", and where $(\cdots)$ and $(\cdots\uparrow)$ mean "statistically not significcant" and "at first not significant but after several months significantly $\uparrow$" respectively.

We can see that $reserves(\uparrow) \rightarrow indpro(\uparrow)$ on the 1st column, and that $reserves(\uparrow) \rightarrow stock(\uparrow) \rightarrow indpro(\uparrow)$ on the 1st and 2nd columns. It should be noticed, however, that $reserves(\uparrow) \rightarrow rex(\downarrow) \not\rightarrow indpro(\uparrow)$ on the 1st and 3rd columns. Therefore, we can conclde that there is a transmission path in QE1 through a stock market, but not through exchange rate.

3.2 Behavior in QE2+α=(2010m11,2012m08)

Since QE2 is too small to construct VAR model, we extend the sampling interval QE2 to $QE2+\alpha = (2010m11, 2011m06) + (2011m07, 2012m08)$ just before starting QE3.

Impulse responses of (1) are depicted in Figure 3, where a solid line implies a calculated impulse response and two dotted lines show 95 % confidence intervals.

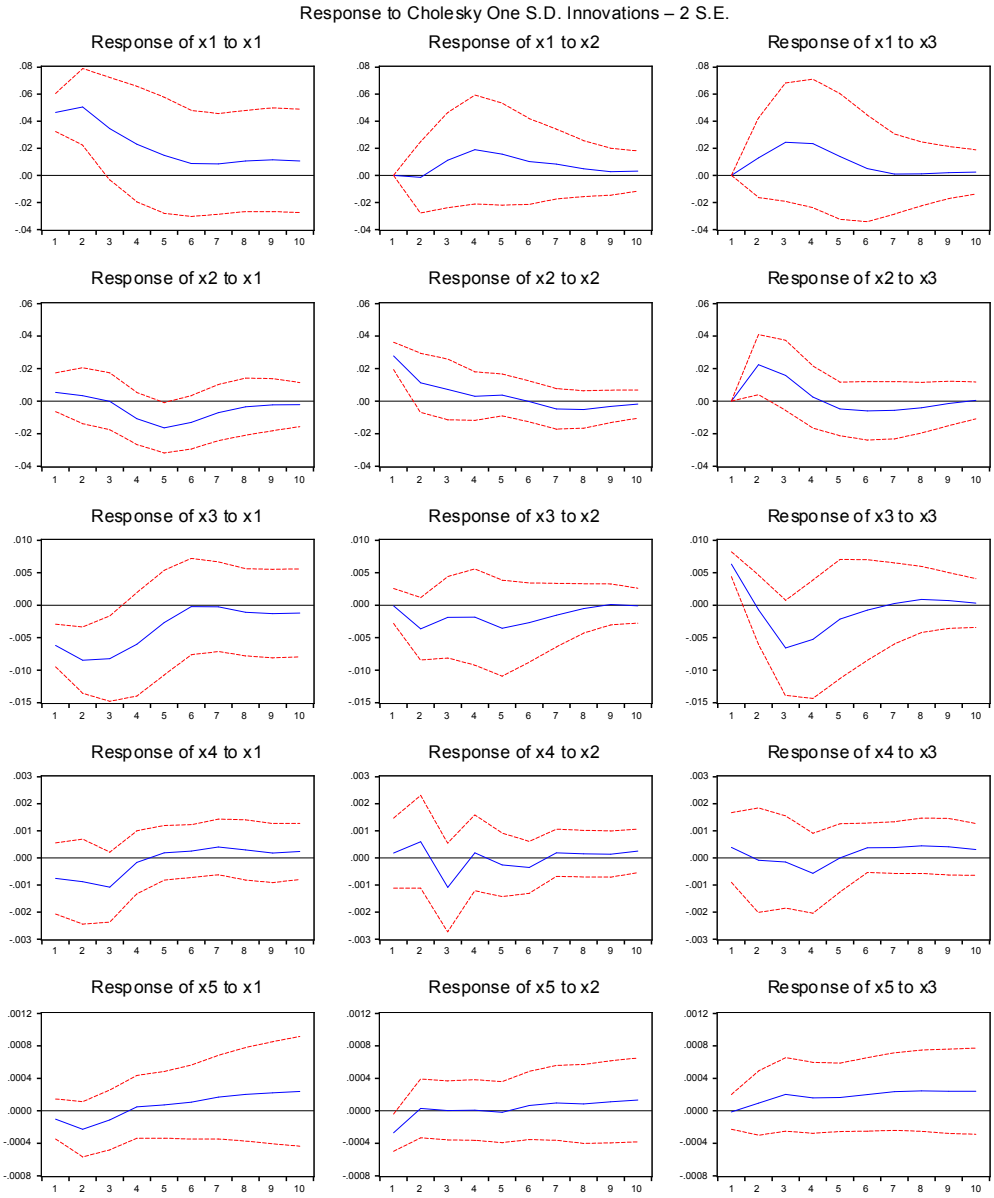


Figure 3. Impulse responses of the system with $x = (x_1, x_2, x_3, x_4, x_5)^{'}$, where in the figure we denote $x_1 = \ln(adjressl/p)$, $x_2 = \ln(sp500/p), x_3 = \ln(twexmmth), x_4 = \ln(indpro), x_5 = \ln(p)$ in $QE2+\alpha = (2010m10, 2012m08)$.

Responses of 5 variables are shown, corresponding to three kinds of impulse shocks; ”reserves“, ”sp500“ and ”exchange rate (twexmmth)“.

1st:	$rsrvs(\uparrow) \Longrightarrow$	$rsrvs(\uparrow)$	$stock(\cdots\downarrow)$
	$rex(\downarrow)$	$indpro(\cdots\downarrow)$	$p(\cdots)$
2nd:	$stock(\uparrow) \Longrightarrow$	$rsrvs(\cdots)$	$stock(\uparrow)$
	$rex(\cdots)$	$indpro(\cdots)$	$p(\downarrow)$
3rd:	$rex(\uparrow) \Longrightarrow$	$rsrvs(\cdots)$	$stock(\uparrow)$
	$rex(\uparrow)$	$indpro(\cdots)$	$p(\cdots)$

In this period of QE2, we cannot say any transmission path from $reserves(\uparrow)$ to $indpro(\uparrow)$.

3.3 Behavior in QE3=(2012m09,2014m10)

Setting sampling interval as QE3, impulse responses of (1) are depicted in Figure.4, where a solid line implies a calculated impulse response and two dotted lines show 95 % confidence intervals. For economy of space, we only show 5 variables responses corresponding to three kinds of impulse shocks; ”reserves“, ”sp500“ and ”exchange rate (twexmmth)“.

1st:	$rsrvs(\uparrow) \Longrightarrow$	$rsrvs(\uparrow)$	$stock(\uparrow)$
	$rex(\downarrow)$	$indpro(\cdots)$	$p(\cdots)$
2nd:	$stock(\uparrow) \Longrightarrow$	$rsrvs(\cdots)$	$stock(\uparrow)$
	$rex(\cdots)$	$indpro(\downarrow)$	$p(\cdots)$
3rd:	$rex(\uparrow) \Longrightarrow$	$rsrvs(\cdots)$	$stock(\cdots)$
	$rex(\uparrow)$	$indpro(\downarrow)$	$p(\downarrow)$

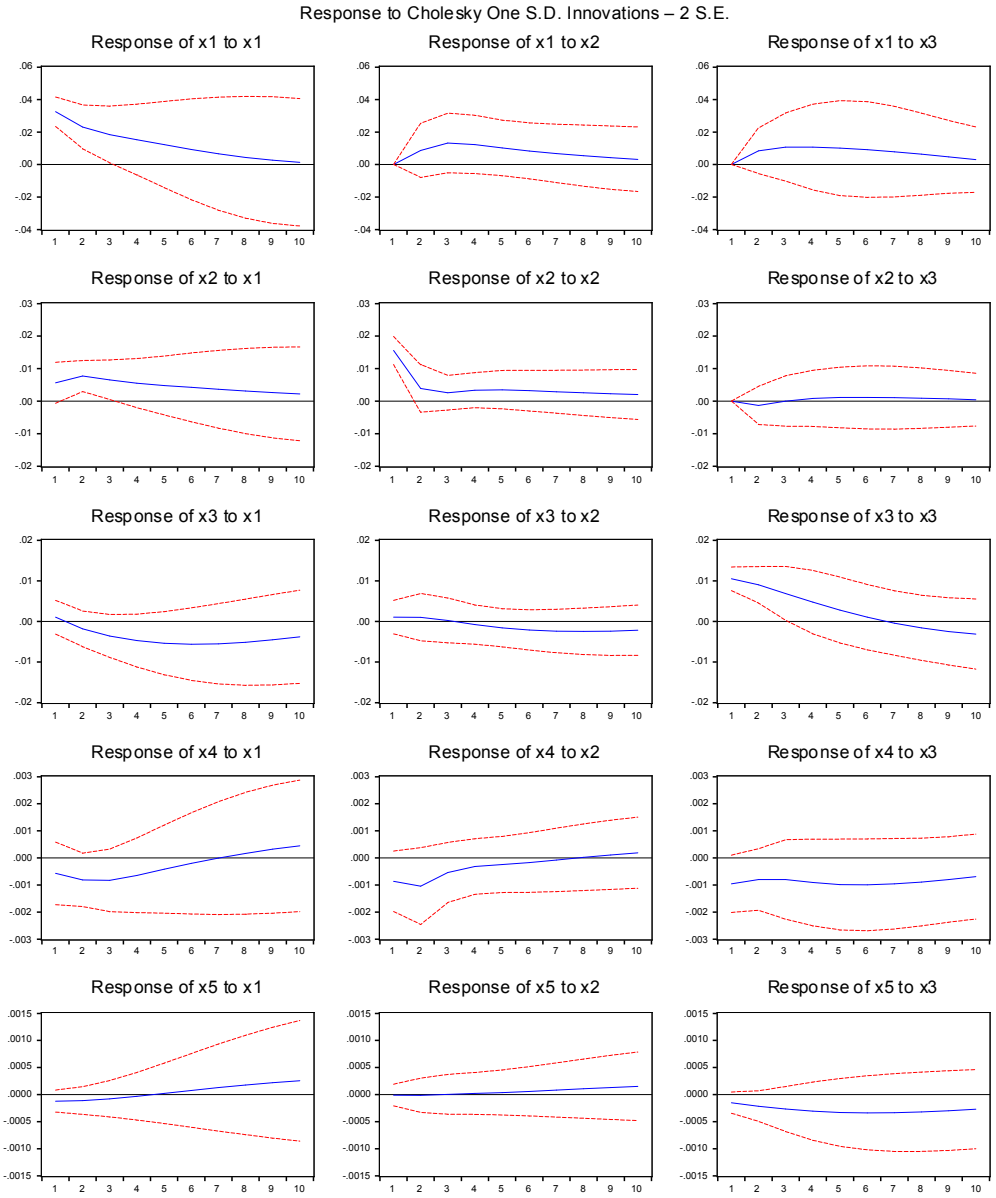


Figure 4. Impulse responses of the system with $x = (x_1, x_2, x_3, x_4, x_5)^{'}$, where in the figure we denote $x_1 = \ln(adjressl/p)$, $x_2 = \ln(sp500/p), x_3 = \ln(twexmmth), x_4 = \ln(indpro), x_5 = \ln(p)$ in QE3=(2012m09,2014m10).

We can see that $reserves(\uparrow) \to rex(\downarrow) \to indpro(\uparrow)$ on the 1st and 3rd columns, while we can see that $reserves(\uparrow) \to stock(\uparrow) \not\to indpro(\uparrow)$ on the 1st and 2nd columns.

Therefore, we can conclude that there is a transmission path in QE3 through exchange rate, but not through a stock market.

4 Transmission Path of Housing Price from Reserve to Economic Activity

In this section, we consider the influence of housing price to the economy. Defining, in (1), $x = (\ln(adjressl/p), \ln(indpro), \ln(cpihossl))'$ and estimating VAR model, we can obtain impulse responses corresponding to the sampling intervals $QE1$, $QE2+\alpha$ and $QE3$. Remark that 5 variables VAR model with $\ln cpilfesl$ replaced by $\ln cpihossl$ in the preceding section gives us a similar result as in this section with 3 variables. For economy of space, we only show the results without figures of impulse responses.

We can see that
(QE1) $reserve(\uparrow) \not\to cpihossl(\uparrow)$,
($QE2+\alpha$) $reserve(\uparrow) \to cpihossl(\uparrow) \to indpro(\uparrow)$,
(QE3) $reserve(\uparrow) \to cpihossl(\uparrow) \to indpro(\uparrow)$.

So, we can conclude that there is a transmission path through housing price in QE2 and QE3, not in QE1.

5 Decomposition of $M2$ into Transaction and Precautionary Money Demands in (1975m10, 2016m03)

In this section, we statistically quantify how much money contributed to the recovery of the economy when the FRB increased reserves. We would decompose the money stock denoted by $m2sl(t)$ into the transaction money and the precautionary money demands.

5.1 Estimation of Precautionary Money Demand

Precautionary money demand will increase when the liquidity concern among the private sector intensify in the depression, while its demand will decrease when the concern dispels in the boom. We use here the $u(t) = napm(t) - 50$, where $napm(t)$ implies "ISM manufacturing: PMI composite index" in order to qualify the unobservable variable, which would affect the precautionary money demand.

Properties of precautionary money demand can be listed as follows:

[Properties of prec. money demand]

- $indpro(\uparrow) \Longrightarrow prec.\ money\ demand(\uparrow)$ as Keynes said.
- $prec.\ money\ demand(\uparrow)$ for future anxiety when economy is in depression.
- prec. money demand is affected by reserves in QEMP.

[Assumption of prec. money demand]

$$\begin{aligned} & prec.\ money\ demand(t) \\ = \ & c_1 * indpro(t) * cpilfesl(t)/1000 \\ + \ & (c_2 * u_n(t) + c_3 * u_p(t)) * \frac{m2sl(t)}{1000} \\ + \ & c_4 * adjressl(t) * dummy_{adjressl}(t) \end{aligned} \tag{2}$$

In the above assumption, the 2nd term on the RHS means that the precautionary money demand is a function of napm, because people try to hold more money when financial anxiety rises, and that the demand may depend on the level of M2. The 3rd term represents effect of the FRB's monetary policies. We take into consideration the policy change by adding the dummy variable. The reserves began to be increased by FRB from September 2008. We have set both of $ff\ rate$ and $adjressl$ as monetary policies, but $ff\ rate$ was not significant, and was deleted in (2). Dummy variable denoted by $dummy_{adjressl}(t)$ in (2) takes value 1 for $t = (2008m09, 2016m03)$ and takes value 0 otherwise.

[Log-likelihood function] The growth rate model of $indpro(t)$ is taken into consideration, and the log-likelihood function of $\Delta\ln(indpro)$ should be maximized with respect to every parameter containing prec. money demand, where in the following equation "$prec.\ money\ demand(t)$" is abbreviated by "$prc.\ mny(t)$".

$$\begin{aligned} & \Delta\ln(indpro(t)) = d_1 * \Delta\ln(indpro(t-1)) \\ + \ & d_2 * \Delta\ln(indpro(t-2)) + d_3 * \Delta\ln(indpro(t-3)) \\ + \ & d_4 * \Delta\ln((m2sl(t-1) - prc.mny(t-1))/p(t-1)) \\ + \ & d_5 * \Delta\ln((m2sl(t-2) - prc.mny(t-2))/p(t-2)) \end{aligned} \tag{3}$$

Table 3 shows estimation results in (2) and (3) .

Table 3. Estimation results of (2) and (3) in $(1975m10, 2016m03)$

	coefficients	*std.error*	*z-statistics*	*prob.*
c_1	143.46	140.57	1.021	0.3074
c_2	-11.82	4.905	-2.410	0.0160
c_3	-4.847	3.256	-1.489	0.1365
c_4	1.003	0.345	2.907	0.0036
d_1	0.118	0.048	2.450	0.0143
d_2	0.176	0.0518	3.399	0.0007
d_3	0.247	0.0389	6.355	0.0000
d_4	0.0425	0.0290	1.468	0.1422
d_5	0.0340	0.0255	1.335	0.1819

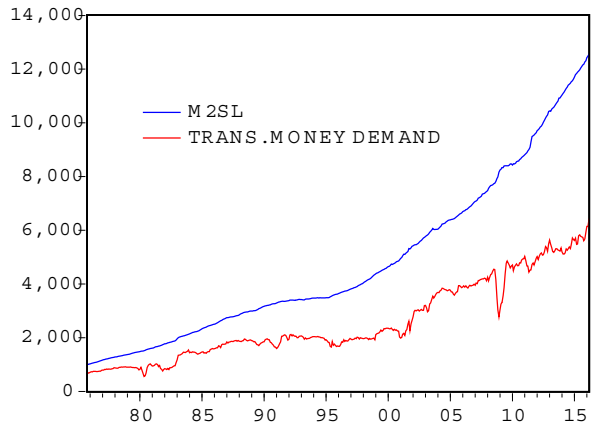


Figure 5. *m2sl* and *trans.money demand* estimated in (1975m10, 2016m03)

Figure 5 shows the nominal money stock *m2sl* and the trans. money demand. The difference "*m2sl* − *trans. money demand*" measures the prec. money demand. We find that the difference begins to expand rapidly around 1995, 2000 and 2008.

Figure 6 depicts the comparison of trans. money demand with prec. one in (2008m09, 2016m03) including QE1,QE2 and QE3. In this figure during the QEMP period, prec. money demand as well as trans. money demand gradually increases except for the beginning of QE1. In the zero interest rates period, there may exist the phenomena of "Liquidity trap" such that easing money by the central bank is only saved without consumption. However, in our estimation, prec. money demand increases while trans. money demand is also increasing. We may insist that the "Liquidity trap" does not exist in QEMP period.

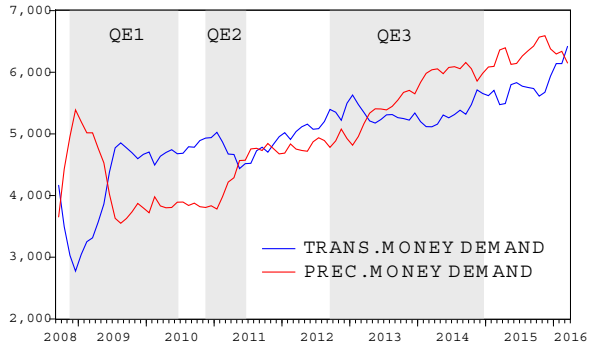


Figure 6. Estimation results of *trans. money demand* and *prec. money demand* in (2008m09, 2016m03)

5.2 The Role of Business Condition $u(t) = napm - 50$ in Transmission Mechanism of QEMP during $QE1, QE2+\alpha$ and $QE3$

We estimate VAR model of $y = (\ln(adjlfesl(t)),\ u(t),\ \ln(trns.\ mny\ dmnd(t)),\ \ln(prc.\ mny\ dmnd(t))^{'}$ with x replaced by y in Eq.(1) . We focus on the role of $u(t)$ in the transmission mechanism of easing monetary policy. Figures 7, 8 and 9 show impulse responses to a one standard deviation shock to four variables in periods of QE1, $QE2+\alpha$ and EQ3 respectively.

In QE1, we can see the following behavior:

1st:	$\ln(adjressl)(\uparrow) \Longrightarrow$	$adjressl(\uparrow)$
	$u(\cdots\uparrow)$	$trns.mny(\downarrow\uparrow)$
	$prc.mny(\uparrow)$	
2nd:	$u(\uparrow) \Longrightarrow$	$\ln(adjressl)(\cdots)$
	$u(\uparrow)$	$\ln(trns.mny)(\uparrow)$
	$ln(prc.mny)(\downarrow)$	
3rd:	$\ln(trns.mny)(\uparrow) \Longrightarrow$	$\ln(adjressl)(\downarrow)$
	$u(\uparrow)$	$\ln(trns.mny)(\uparrow)$
	$ln(prc.mny)(\downarrow)$	
4th:	$\ln(prc.mny)(\uparrow) \Longrightarrow$	$adjressl(\uparrow)$
	$u(\cdots\uparrow)$	$trns.mny(\cdots\uparrow)$
	$prc.mny(\uparrow)$	

Summarizing the behavior in QE1, we can say that a quantitative monetary easing has a positive effect on USA's economy. On the 1st column of Figure 7, we can see first $reserve(\uparrow) \rightarrow prec.demand(\uparrow)$. At the same time, *trans.demand* changes from downward to upward during several months: $trans.demand(\downarrow \cdots \uparrow)$. u rises along with *trans.demand* on the 1st column, while on the 2nd column, $u(\uparrow) \rightarrow trans.demand(\uparrow)$.

In $QE2+\alpha$, we can't see the path from reserve to trans. demand in Figure 8.

In QE3, on the 1st and 2nd columns of Figure 9, we can see
$reserve(\uparrow) \not\rightarrow trans.demand(\uparrow)$,
$reserve(\uparrow) \rightarrow u(\uparrow)$,
$u(\uparrow) \rightarrow trans.demand(\uparrow)$.
Thus, we can say that efficency of QEs is, in order, given by $QE1 > QE3 > QE2$.

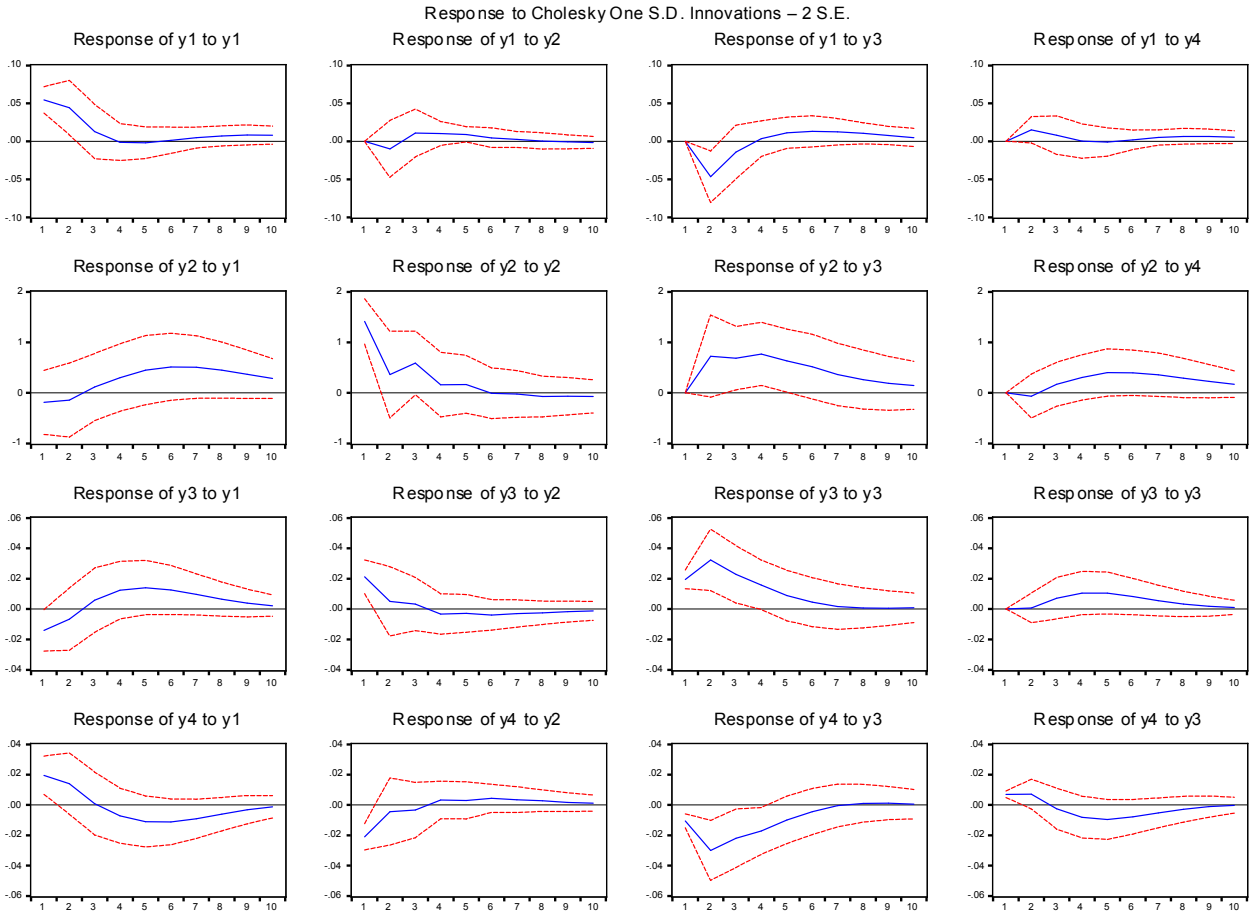


Figure 7. Impulse responses of the system with $y = (y_1, y_2, y_3, y_4,)'$, where in the figure we denote $y_1 = \ln(adjressl/p), y_2 = u(t)$, $y_3 = \ln(trans.\ money\ demand(t)), y_4 = \ln(prec.\ money\ demand(t))$ in QE1.

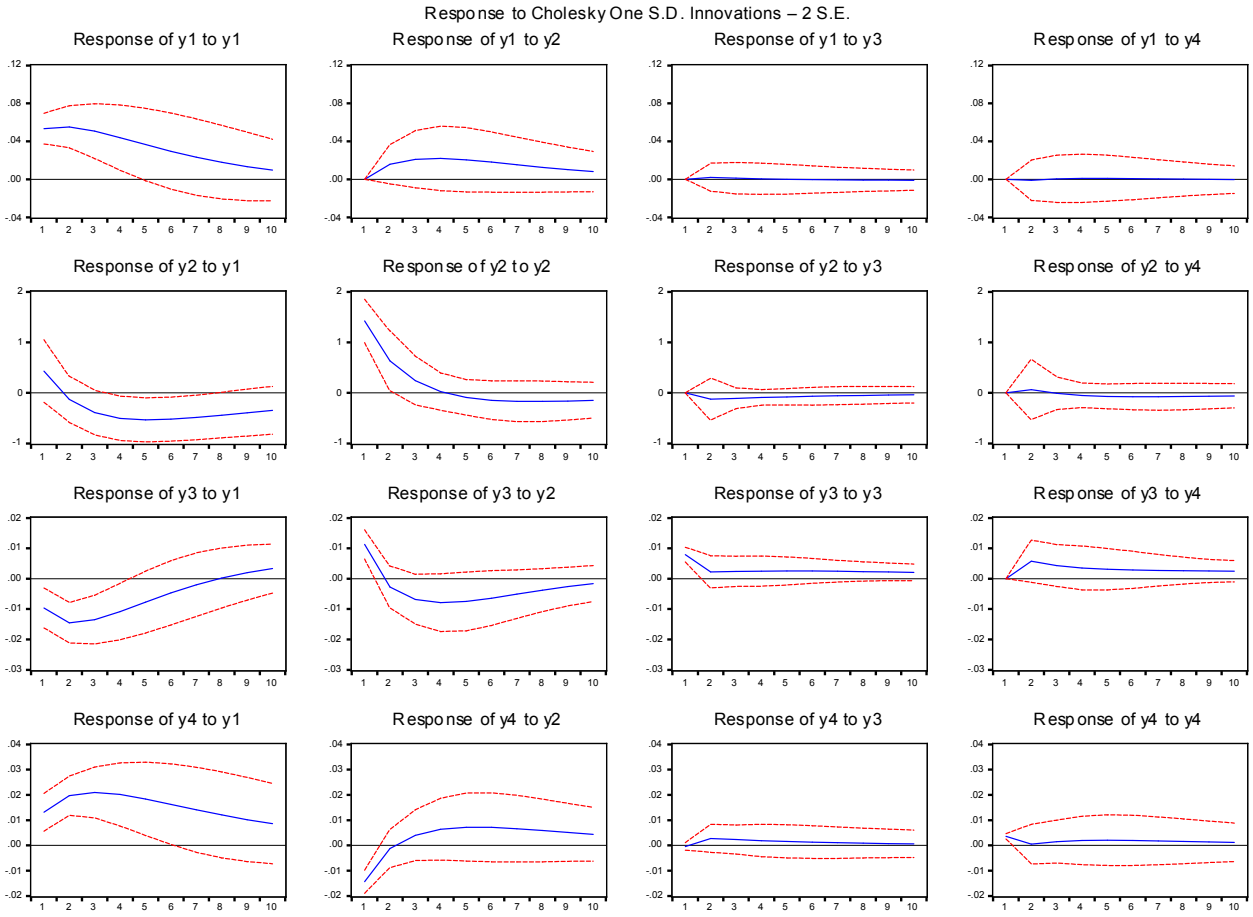


Figure 8. Impulse responses of the system with $y = (y_1, y_2, y_3, y_4,)^{'}$, where in the figure we denote $y_1 = \ln(adjressl/p), y_2 = u(t)$, $y_3 = \ln(trans.\ money\ demand(t)), y_4 = \ln(prec.\ money\ demand(t))$ in $QE2 + \alpha$.

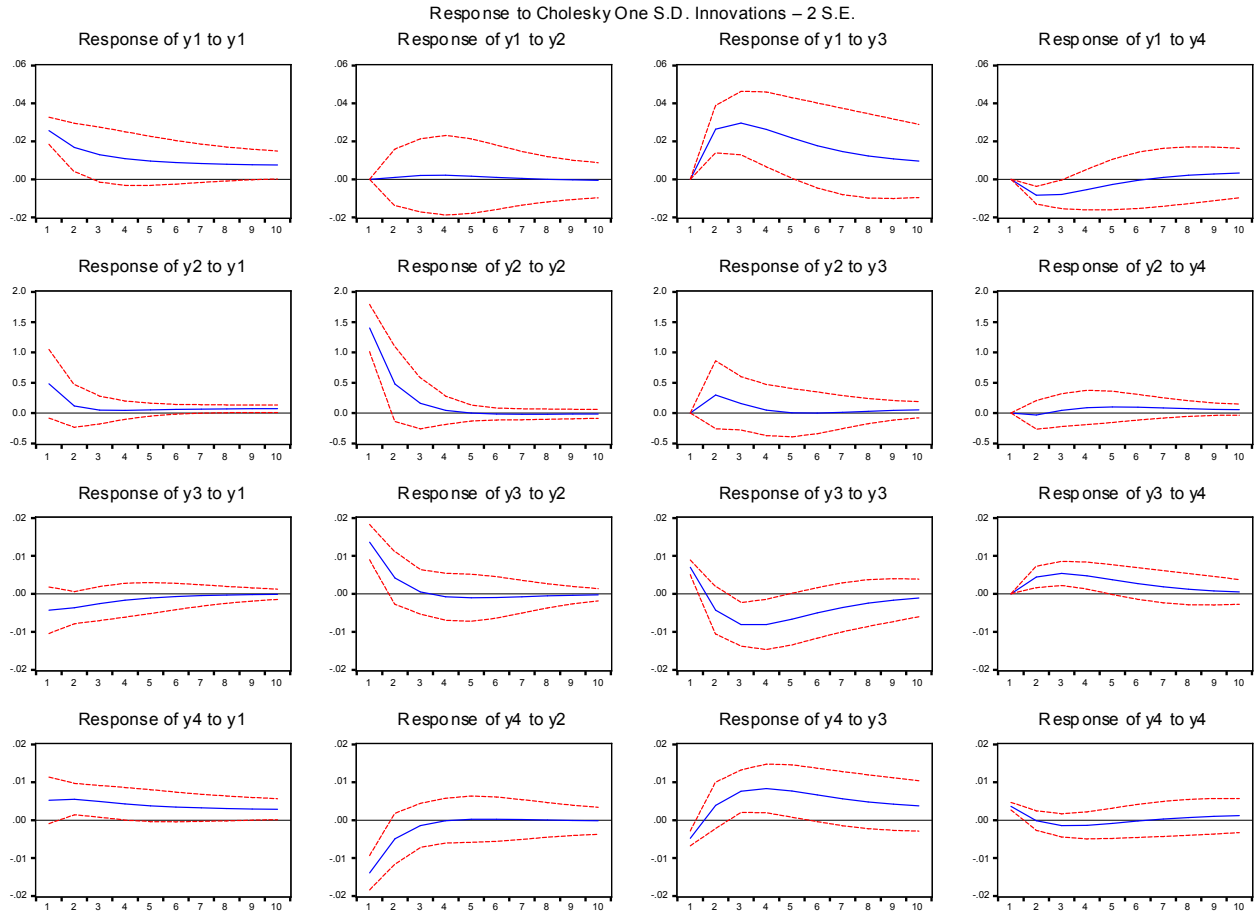


Figure 9. Impulse responses of the system with $y = (y_1, y_2, y_3, y_4,)$', where in the figure we denote $y_1 = \ln(adjressl/p), y_2 = u(t)$, $y_3 = \ln(trans.\ money\ demand(t)), y_4 = \ln(prec.\ money\ demand(t))$ in QE3.

6 Conclusions

We investigated efficiency of QE1, QE2 and QE3 in USA in two ways. First, usual VAR models were constructed and we can see that QE1 is effective through a transmission path of a stock market, QE2 and QE3 are effective through housing price path, and that QE3 is effective through an exchange rate path. Secondly, decomposing M2 into precautionary and transaction money demands, we can estimate precautionary money demand as a function on "industrial production", business condition and reserves. By investigating relationship among reserves, business condition, transaction and money demands, we see that QE1 is most effective and that QE3 is effective and that QE2 is not so effective. .

References

B.S. Bernanke. The crisis and the policy response. Technical report, London School of Economics, 2009.

S. Bhattarai, A. Chatterjee, and W.Y. Park. Effects of US quantitative easing on emerging market economies. *Globalization and Monetary Policy Institute, Working Paper*, (255), 2015.

E. Engen and D. Reifschneider. The macroeconomic effects of the Federal Reserve's unconventional monetary policies. *Finance and Economics Discussion Series*, (2015-005), 2015.

B.W. Fawley and C.J. Neely. Four stories of quantitative easing. *Federal Reserve Bank of St. Louis REVIEW*, 95(1):51–88, 2013.

M. Fratzscher, M.L. Duca, and R. Straub. On the international spillovers of US quantitative easing. *European Central Bank Working Paper Series*, (1557), 2013.

J. Gagnon, M. Raskin, J. Remache, and B. Sack. Large-scale asset purchases by the Federal Reserve: Did they work? *Economic Policy Review*, 2011.

S.G. Hall, P.A.V.B. Swamy, and G.S. Taylas. Milton Friedman, the demand for money, and the ECB's monetary policy strategy. *Federal Reserve Bank of St. Louis REVIEW*, 2012.

Y. Honda, Y. Kuroki, and M. Tachibana. An injection of base money at zero interest rates: empirical evidence from the Japanese experience 2001-2006. *Osaka University, Discussion Papers in Economics and Business*, 7(8), 2007.

Y. Morita and S. Miyagawa. Efficiency of quantitative easing in Japan during (2001, 2006) through estimation of precautionary money demand-II. In *Proceedings of SSS15 (the ISCIE International Symposium on Stochastic Systems Theory and Its Applications)*, 2016.

Y. Sawada. The effectiveness of unconventional monetary policy in USA. In *Proceedings of SSS13 (the ISCIE International Symposium on Stochastic Systems Theory and Its Applications)*, 2014.

Y. Wen. When and how to exit quantitative easing? *Federal Reserve Bankorita of St. Louis Review*, 96(3):243–265, 2014.

Riser of Dual Fluidized Bed Gasification Reactor: Investigation of Combustion Reactions

Rajan Kumar Thapa Britt M. E. Moldestad

Department of Process, Energy and Environmental Technology
University College of Southeast Norway, Porsgrunn, Norway
Rajan.k.thapa@usn.no

Abstract

The riser of a dual fluidized bed gasification reactor heats bed materials by burning residual char particles coming from gasification part of the reactor. A validated Computational Particle Fluid Dynamic (CPFD) model is applied to simulate combustion of char particles in a riser of dual fluidized bed gasification reactor in a demonstration plant with 8 MW fuel capacity. The plant is located in Güssing, Austria. The three-dimensional model is used investigate combustion reaction as a function of the bottom, primary and secondary air feed rates. The results show there is a still possibility to improve combustion reaction by optimizing air feed rates, which can maximize the bed material temperature without increasing additional char particles feed.

Keywords: dual fluidized bed, gasification, biomass, combustion, riser

1 Introduction

Biomass gasification is one of the promising technologies for combined heat and power production and synthesis processes leading to the production of liquid biofuels. Biomass has two major advantages: it is carbon dioxide neutral and homogeneously and locally available all over the world (Asadullah, 2014). There are various types of gasification technologies such as fix bed, moving the bed and fluidized bed (Basu, 2013). Dual fluidized bed steam gasification is one of the latest technologies among them (Göransson et al., 2011; Hofbauer et al., 2001).

Steam gasification in a dual fluidized bed reactor is a complex thermochemical process by which biomass is converted to a mixture of combustible and non-combustible gasses and other minor components. The combustible gasses are called producer gas. The major components of the producer gas are carbon monoxide, hydrogen, and methane. The non-combustible gasses are carbon dioxide and water vapor (Hofbauer et al., 1997, 2002a,b). The principle of the dual fluidized bed gasification reactor is shown in Figure 1. The reactor consists of two parts where one is a bubbling fluidized bed gasification reactor, and the other is a circulating fluidized bed combustion reactor. The gasification reactions are endothermic, and heat required for the reactions is supplied by circulating hot bed materials from the combustion reactor (Kern et al., 2013). The bed material can be sand or olivine particles. The primary purpose of the bed material is to transfer heat from the combustion reactor to the gasification reactor. The bed material is heated in the combustion reactor by burning char particles. The char particles, are residual char after the gasification process, which is transported from the gasification reactor to the combustion reactor along with the bed materials.

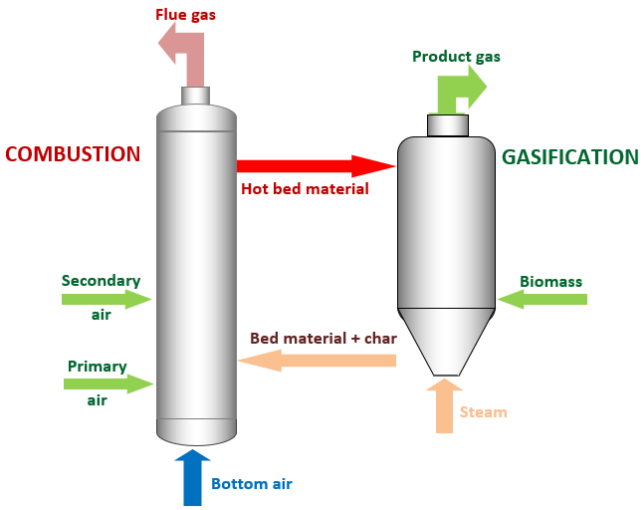


Figure 1. Principle of dual fluidized bed gasification system.

To optimize the performance of the reactor, gas-particle flow and combustion reaction should be optimized. It means that the reactor should have a minimum fluctuation of bed material temperature, steady transport of hot bed material to the gasification reactor and optimum gas flow rates (Snider and Banerjee, 2010). The temperature of bed material depends on the combustion reaction. The combustion reactions are dependent on the amount of oxygen supply. In the combustion reactor, the oxygen for the combustion reaction is provided from the air which is fed to the reactor from three position as the bottom, primary and secondary air.

The temperature of the bed material is self-controlled by the reactor as a whole. If the temperature of circulated bed material is low, the gasification reaction rates

become slower leaving more unreacted residual char particles to circulate to the combustion reactor. The more char particles circulated to the combustion reactor, the more the temperature of the bed materials is increased and vice versa (Pfeifer et al., 2011; Wang and Chen, 2013). The self-stabilization process requires two essential conditions to be fulfilled. Firstly, the initial temperature of the bed materials should be maintained near to the reaction temperature (about 850°C) which is achieved by burning a part of the producer gas to start up the reactor. Secondly, it is important to maintain constant bed material circulation rate which depends on the air feed rates and air feed positions in the reactor. Previous studies performed by the authors showed that the optimum ratio of bed materials to biomass feed rate was 25-30 for Güssing plant (Thapa and Halvorsen, 2014).

Preheated air is fed to the combustion reactor from the bottom and two positions along the height of the reactor as primary and secondary air. The air supplies necessary oxygen for combustion reaction and simultaneously serves as fluidizing gas to transport heated bed materials to the gasification reactor. The feed air is preheated to achieve better combustion of char in the reactor.

It is not well understood, whether all residual char particles coming from the gasification reactors are totally combusted during the flow along the riser of the combustion reactor. It is because the gas velocity in the combustion reactor is high and the residence time may not be long enough for all char particles to undergo complete combustion. Moreover, the air feed should ensure sufficient oxygen to burn all char particles.

The fluid dynamics of the combustion reactor requires the lower part of the reactor to be in bubbling fluidization regime, which means that the bottom air feed velocity should be lower. If the velocity is very high, a part of flue gas can pass to the gasification reactor making the product gas diluted by nitrogen and carbon dioxide contained in the flue gas, which is undesirable. The middle and upper parts are in the fast fluidization regime (Kaushal et al., 2008a). Moreover, the flow properties in the reactor vary with height because of the three different feed position of the bottom, primary and secondary air. The flow parameters are different at different temperature due to density and viscosity variation of the fluidizing gas.

A three-dimensional CPFD model is developed to study and optimize fluid dynamic properties and reaction kinetics in the combustion reactor. The gas-particle flow is investigated in high-temperature fluid flow with char combustion. The model is simulated using the commercial Computational Particle Fluid Dynamic (CPFD) software Barracuda VR. 15. The effect of char combustion is studied with varying flow rates of the bottom, primary and secondary airflow rates. The aim of the series of simulation is to investigate char combustion in the reactor and flow rates of the bottom, primary and secondary airflow rates. All parameters used in the simulations are based on the combustion riser in the biomass gasification plant in Güssing, Austria.

2 Mathematical Model

In this work, a Computational Particle Fluid Dynamic (CPFD) model is applied to simulate the gas-solid flow with heat transfer and chemical reactions. The CPFD numerical methodology incorporates multi-phase-particle-in-cell (MP-PIC) method (Andrews and O'Rourke, 1996; Snider, 2001). The gas phase is solved using Eulerian grid and the particles are modeled as Lagrangian computational particles. Gas and particle momentum equations are solved in three dimensions. The fluid is described by the Navier-Stokes equation with strong coupling to the discrete particles. The particle momentum follows the MP-PIC description which is a Lagrangian description of particle motions described by ordinary differential equations coupling with the fluid (Snider and Banerjee, 2010).

In the CPFD numerical method, actual particles are grouped into computational particles, each containing a number of particles with identical densities, volume and velocities located at a specific position. The computational particle is a numerical approximation similar to the numerical control volume where a spatial region has a single property for the fluid. With these computational particles, large commercial systems containing billions of particles can be simulated using millions of computational particles. This possibility of the CPFD numerical method is used in this work to simulate the riser part of the large scale dual fluidized bed steam gasification plant.

Governing equations

The volume averaged fluid mass and momentum equations are:

$$\frac{\partial(\varepsilon_g \rho_g \mathbf{u}_g)}{\partial t} + \nabla(\varepsilon_g \rho_g \mathbf{u}_g \mathbf{u}_g) = \varepsilon_g \nabla p - \mathbf{F} + \varepsilon_g \rho_g \mathbf{g} + \varepsilon_g \tau_g \quad (1)$$

where ε_g, ρ_g, and $\mathbf{u}_g$ are gas volume fraction, density and velocity respectively, p is gas pressure, $\mathbf{g}$ is the acceleration due to gravity, $\mathbf{F}$ is the rate of momentum exchange per unit volume between the gas and solid phase and τ_g is stress tensor which can be expressed in index notation as:

$$\tau_{g,ij} = \mu \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) - \frac{2}{3} \mu \delta_{ij} \frac{\partial u_k}{\partial x_k} \quad (2)$$

where μ is shear viscosity. The shear viscosity is the sum of laminar shear viscosity and turbulence viscosity based on the Smagorinsky turbulence model. In the model, large eddies are directly calculated. The unresolved sub-grid turbulence is modeled by using eddy viscosity. The turbulence viscosity is given as:

$$\mu_t = C \rho_g \Delta^2 \sqrt{\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)^2} \quad (3)$$

where C is sub-grid eddy coefficient and known as Smagorinsky coefficient.

MP-PIC method calculates the particle phase dynamics using the particle distribution function (PDF), f_p. A transport equation is solved for the PDF. The transport equation for f_p is given by [14] and is expressed as:

$$\frac{d_{f_p}}{dt}+\frac{\partial(f_p u_p)}{dx}+\frac{\partial(f_p A_p)}{du}=\frac{f_D-f_p}{\tau_D} \tag{4}$$

where u_p is particle velocity, f_D is the particle distribution function for the local mass averaged particle velocity and τ_D is the collision damping time. A_p is the particle acceleration which is given by:

$$A_p=\frac{\partial u_p}{dt}=D_p\,(u_g-u_p)-\frac{1}{\rho_p}\nabla p_g+g-\frac{1}{\varepsilon_p\rho_p}\nabla\tau_g+g+F_p \tag{5}$$

In the equation above, ε_p is particle volume fraction, ρ_p is particle density, p_g is gas pressure, τ_p is contact normal stress. More details about the τ_p can be found in (O'Rourke and Snider, 2010). F_p is the particle friction per unit mass, related to the relative particle motion and becomes important at very low particle flow at near closed packed bed (Snider, 2007) and D_p is the drag function. The Wen-Yu drag model is implemented in this work (Wen, 1966).

$$D_p=C_D\frac{3}{8}\frac{\rho_g}{\rho_p}\frac{|u_g-u_p|\varepsilon_g^{-2.65}}{r_p} \tag{6}$$

where

$$D_p=C_D\begin{cases}\frac{24}{Re}\left(1+0.15\,Re^{0.678}\right), & Re<1000\\ 0.44, & Re\geq 1000\end{cases} \tag{7}$$

$$Re=\rho_g\frac{|u_g-u_p|r_p}{\mu_g}\quad\text{and}\quad r_p=\left(\frac{m}{\frac{4}{3}\pi\rho_p}\right)^{\left(\frac{1}{3}\right)} \tag{8}$$

3 Model Parameters and Geometry

The dimensions of the reactor are the same as the combustion reactor in the biomass gasification plant in Güssing, Austria. The basic dimensions of the combustion part of the reactor are shown in Table 1.

Table 1. Reactor Dimensions.

Dimensions	*Units*	*Value*
Diameter	m	0.66
Height	m	12
Primary air inlet	m	1.5
Secondary air inlet	m	3.5

In the plant, the gasification and combustion occur simultaneously in the bubbling fluidized bed gasification reactor and in circulating fluidized bed combustion reactor. The whole reactor is a combination of these two. However, the aim of current study is only the combustion part of the reactor. Therefore, the combustion part of the reactor is separated in the model replacing circulation of the bed materials by the inlet and outlet boundaries. Dotted lines in Figure 2 show the control volume of model geometry.

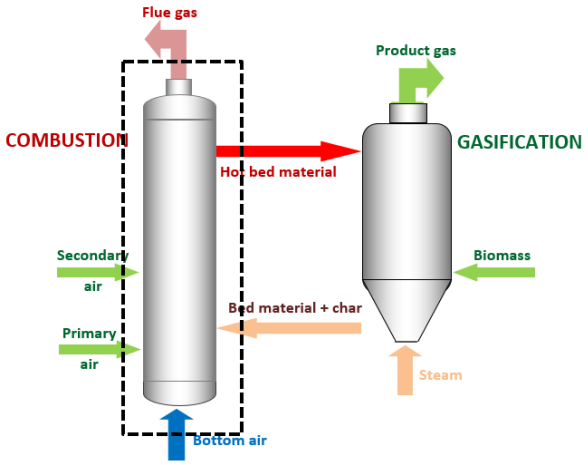


Figure 2. Control volume of model geometry.

In the CPFD computational model, grid generating and solving flow and reactions in a rectangular geometry is better than in a circular geometry. For this reason, the circular diameter of the geometry is converted into rectangular with the equivalent cross-sectional area. The geometry used in the CPFD model with all boundary conditions and grids are shown in Figure 3.

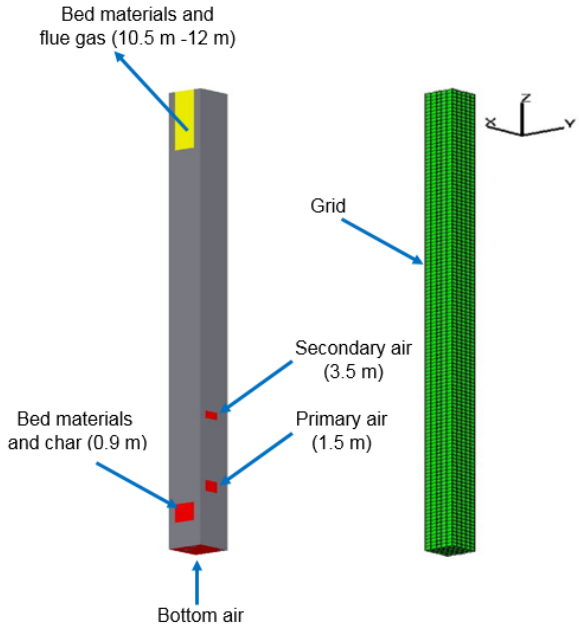


Figure 3. Grid and boundary conditions of the riser.

The combustion reactor uses air as a fluidizing agent as well as an oxidizing agent. The gas-particle flow involves a complex mixture of air, flue gas, olivine particles

and char. The properties of gas and particles used in the simulation are given in Table 2.

Table 2. Properties of Solids and Gases.

Properties	*Units*	*Value*
Olivine particle size	μm	200-800
Olivine density	kg/m^3	2960
Char particles size	mm	1-5
Char density	kg/m^3	200
Air density	kg/m^3	0.27-1.06
Air temperature inlet	K	333-1300
Air viscosity	$Pa \cdot s$	$[1.98\text{-}4.9]\cdot 10^{-5}$

In the combustion reactor, olivine particles and char have a wide range of size distribution. The size distribution of char and bed materials are presented in Figure 4(a) and 4(b) respectively.

The particle sizes expressed in the figure are the radius of particles. It is because Barracuda uses radius to represent particle size. The combustion reactions and their reaction kinetics involved in the CPFD model is given in Table 3 (Kaushal et al., 2008a).

Chemical reactions can affect gas flow rates, gas compositions, particle sizes and particle densities. The reactions, on the other hand, can be affected by temperature, gas-particle mixing and gas feed positions and feed rates in the reactor. There is strong interdependence between reaction chemistry and particle-fluid dynamics. Therefore, it is important to take into account the change in flow behavior due to the combustion reaction in the riser.

The volume average chemistry is used in the current model. The gas volume of each cell in the grid acts as control volume for the reaction calculations. In volume average chemistry, each control volume is the gas volume in a cell. The reactions are written in stoichiometric form. Temperature, pressure and solid dependence are entered as rate coefficients.

The particle dependency term is the radius of the particle. The temperature is the average temperature of the particle and the bulk fluid.

It is assumed that the particle temperature is constant within the particle.

4 Results and Discussions

To simulate the riser of the Güssing plant and to investigate the combustion reaction, the first series of simulations were run with the gas and fuel feed parameters as reported from the biomass gasification plant in Güssing (Kaushal et al., 2008a). The parameters are presented in Table 4.

The table shows the flow rate of the bottom, primary and secondary air with different temperatures. The reason for feeding the air at various position and temperatures is to control the combustion process and bed material circulation rate. The mixture of bed materials and residual char is fed to the combustion reactor at a temperature of 1073 K. The char is combusted in the reactor and it gradually heats the bed materials while moving up along the height of the riser. Therefore, the particle temperature is gradually increased as they move towards the top of the reactor. The temperature of the gas and particles are measured from the bottom and every one-meter height of the reactor. The simulated fluid and particle temperature along the height of the reactor is shown in Figure 5.

The figure indicates that the fluid and particle temperature are changing with height in the riser. At every position in the reactor, the fluid temperature is higher than particle temperature. It indicates that the heat transfer between particles and fluid is not sufficient due to high velocity and low residence time. The fluid temperature deviates sharply from particle temperature above the height of 2 m. This point is just above the primary air feed point. Preheated primary air significantly improves the char combustion process in the reactor and increases the temperature. However, after about 8 m height, the fluid and particle temperature do not increase indicating that char combustion does not occur above this height of the bed.

The char particle volume fraction and the mole fraction of the major components of the flue gas is presented in Figure 6. The figure shows that the residual char particles fed to the combustion reactor are not completely burned. The existence of the small amount of char particles at the top of the reactor suggests that the char particles are recirculated back to the gasification reactor. As expected, the mole fraction of CO_2 increasing from the bottom to the top while the oxygen mole fraction is decreasing. The oxygen mole fraction is almost zero at a height about 7 m implying that there is an insufficient amount of oxygen to obtain total combustion of char particles. The small amount of H2 at the top of reactor makes that fact clearer.

The *CO* concentration is a strong function of the available oxygen. Hence, the primary air flow which is the biggest among the all air flow rates influences the *CO* concentration most (Kaushal et al., 2008b).

A series of simulation were run by gradually increasing the bottom airflow rate. The results are presented in Figure 7.

An increasing amount of bottom air increases the gas and particle temperature. The maximum particle temperature is increased from about 1150K to about 1220K in the first series of simulations (Figure 5). The increase in temperature without increasing the mass flow rate of the char particles indicates that not all the char particles are burned. In other words, there is still some uncombusted char particles passing through the combustion reactor to the gasification reactor. However, increasing the bottom air increases the fluid temperature much more than the particle temperature. It may be due to decreased residence time for the particles in the combustion reactor. Increased bottom air flow rate increases the fluidization velocity making particles be transported faster. In addition, the high feed rate of bottom air is not desirable due to the risk of flue gas leakage to the gasification reactor.

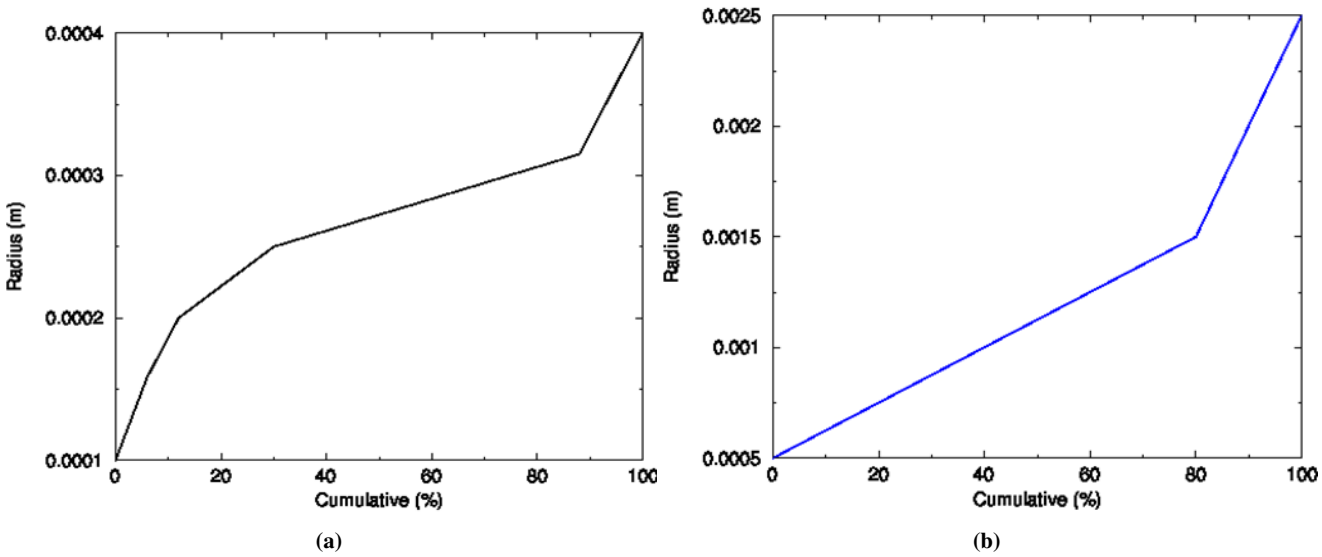


Figure 4. Particle size distribution (a) Olivine (b) Char.

Table 3. Reactions and Reaction Kinetics.

Reactions	*Reaction Rate*	K_1
$C+O_2=CO+CO_2$	$R_1=K_1[O_2]^{0.5}$	$8.56\cdot10^{-2}exp(\frac{-2237}{T})$
$C+H_2O=CO+H_2$	$R_2=K_1[H_2O]^{0.57}$	$2.62\cdot10^{8}exp(\frac{-237000}{T})$
$H_2+\frac{1}{2}O_2=H_2O$	$R_3=K_1[H_2]^{1.5}[O_2]$	$1.63\cdot10^{9}T^{3/2}exp(\frac{-3420}{T})$
$CO+\frac{1}{2}O_2=CO_2$	$R_4=K_1[H_2]^{1.5}[O_2]^{0.5}[H_2O]^{0.5}$	$3.25\cdot10^{7}exp(\frac{-15098}{T})$
$CO+H_2O=H_2+CO_2$	$R_5=K_1[CO][H_2O]$	$0.03\ exp(\frac{-7249}{T})$

Table 4. Flow Parameters - Gussing Plant.

Parameters	*Feed Rate* $[Nm^3/h]$	*Temperature [K]*
Bottom air	720	333
Primary air	2880	673
Secondary air	869	860
Bed materials	37 [kg/s]	-

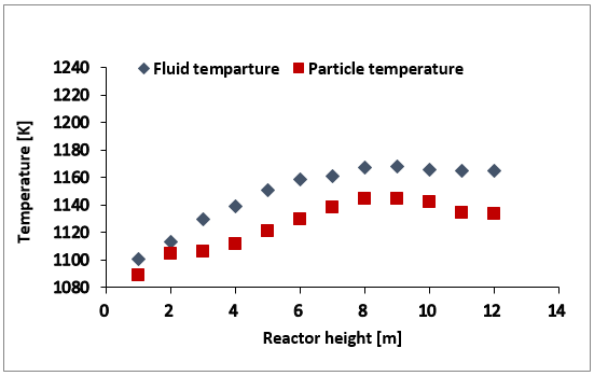


Figure 5. Fluid and particle temperature along the height of the reactor.

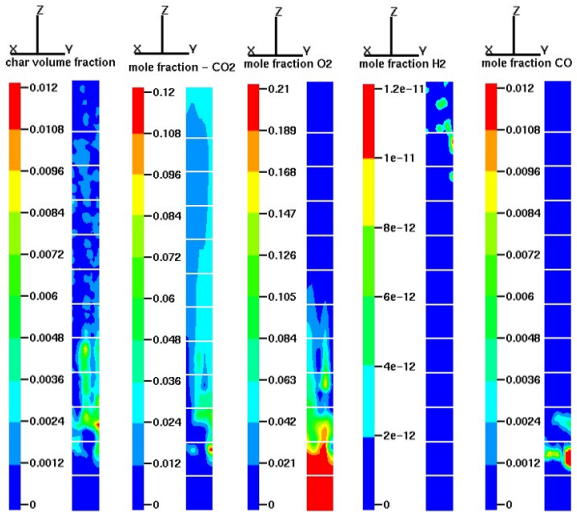


Figure 6. Particles and gas fractions.

A series of simulations were run to investigate the change in fluid-particle temperature with increasing primary air flow rate. The results presented in Figure 8 shows that the particle temperature is again increasing with the increasing primary air flow rate. Moreover, the difference between the fluid and particle temperature is also de-

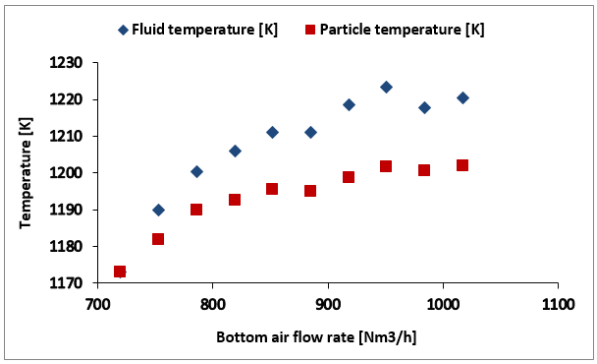


Figure 7. Fluid and particle temperature at a height of 10 m with increasing bottom air flow rate.

creased in this case.

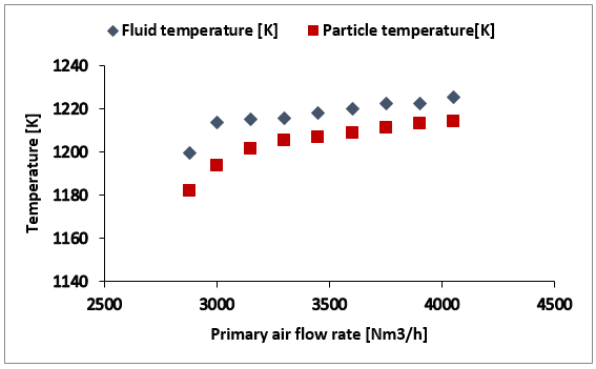


Figure 8. Fluid-particle temperature with increasing primary air flow rate.

The results of the simulation series with increasing secondary air flow rate also show that there is a significant change in fluid-particle temperature as shown in Figure 9.

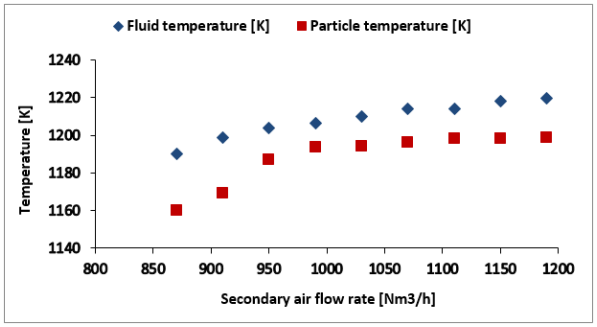


Figure 9. Fluid particle temperature with increasing secondary air flow rate.

The overall particle temperature distribution along the combustion reactor is presented in Figure 10.

The particle temperature distribution at increasing air flow rates of the bottom, primary and secondary air are compared with the particle temperature at the air flow rate used at the Güssing plant. The results show that increasing primary and secondary air flow rates result in higher and more uniform particle temperature distribution in the

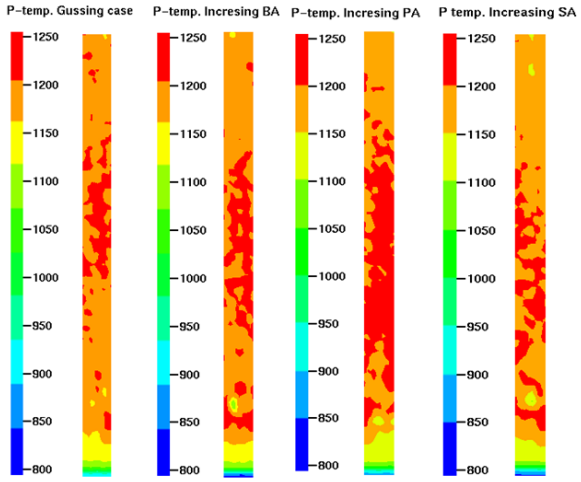


Figure 10. Snapshot of particle temperature distribution along the combustion reactor at 400s of simulation time.

reactor.

5 Conclusions

A 3D CPFD model is developed to investigate the effect of bottom, primary and secondary air flow rates in the combustion process and bed material temperature in the riser of a dual fluidized bed biomass gasification reactor. The simulated reactor is located at 8 MW fuel, biomass gasification plant in Güssing, Austria. Series of simulations were run using commercial CPFD software Barracuda V15. The first series of simulation were performed with all parameters as in the Güssing plant. The results show that some un-combusted char particles are passing through the reactor. Three series of simulations carried out with increasing bottom, primary and secondary air feed rate. All the cases show increasing temperature with increasing air feed rates at a constant feed rate of char particles. The results of the simulations also indicate that increase in primary air flow rate results in the highest particle temperature with more uniform temperature distribution along the reactor than the cases with increasing bottom and secondary air feed rate. The results show that there is still a possibility for optimization of air flow rates for combustion reaction in the riser.

References

Michael J. Andrews and Peter J. O'Rourke. The multiphase particle-in-cell (MP-PIC) method for dense particulate flows. *International Journal of Multiphase Flow*, 22(2):379–402, 1996. doi:10.1016/0301-9322(95)00072-0.

Mohammad Asadullah. Barriers of commercial power generation using biomass gasification gas: a review. *Renewable and Sustainable Energy Reviews*, 29:201–215, 2014. doi:10.1016/j.rser.2013.08.074.

Prabir Basu. *Biomass gasification, pyrolysis and torrefac-*

tion: practical design and theory. Academic press, 2013.

Kristina Göransson, Ulf Söderlind, Jie He, and Wennan Zhang. Review of syngas production via biomass DFBGs. *Renewable and Sustainable Energy Reviews*, 15 (1):482–492, 2011. doi:10.1016/j.rser.2010.09.032.

Hermann Hofbauer, Rauch Reinhard, Bosch Klaus, Koch Reinhard, and Aichernig Christian. Biomass CHP plant güssing–a success story. *Available online: http://members. aon. at/biomasse/strassbourg. pdf (accessed on 14 December 2016)*, 2001.

Hermann Hofbauer, Rauch Reinhard, Bosch Klaus, Koch Reinhard, and Aichernig Christian. Biomass CHP plant güssing–a success story. *Expert Meeting on Pyrolysis and Gasification of Biomass and Waste, Strasbourg, France*, 2002b.

Hermann Hofbauer, Reinhard Rauch, Gerhard Löffler, Sebastian Kaiser, E Fercher and H Tremmel. *Six years experience with the FICFB-gasification process*. 12th European Conference and Technology Exhibition on Biomass, Energy, Industry and Climate Protection, Amsterdam, 2002a.

Hermann Hofbauer, Günter Veronik, Thomas Fleck, Reinhard Rauch, Herbert Mackinger, and Erich Fercher. *The FICFB-gasification process*. Springer, 1997. doi:10.1007/978-94-009-1559-682.

Priyanka Kaushal, Tobias Pröll, and Hermann Hofbauer. Model for biomass char combustion in the riser of a dual fluidized bed gasification unit: Part i model development and sensitivity analysis. *Fuel Processing Technology*, 89(7):651–659, 2008a. doi:10.1016/j.fuproc.2007.12.010.

Priyanka Kaushal, Tobias Pröll, and Hermann Hofbauer. Model for biomass char combustion in the riser of a dual fluidized bed gasification unit: Part ii model validation and parameter variation. *Fuel Processing Technology*, 89(7):660–666, 2008b. doi:10.1016/j.fuproc.2007.12.009.

Stefan Kern, Christoph Pfeifer, and Hermann Hofbauer. Gasification of lignite in a dual fluidized bed gasifier–influence of bed material particle size and the amount of steam. *Fuel processing technology*, 111:1–13, 2013. doi:10.1016/j.fuproc.2013.01.014.

Peter J. O'Rourke and Dale M. Snider. An improved collision damping time for MP-PIC calculations of dense particle flows with applications to polydisperse sed-imenting beds and colliding particle jets. *Chemical Engineering Science*, 65(22):6014–6028, 2010. doi:10.1016/j.ces.2010.08.032.

Christoph Pfeifer, Stefan Koppatz, and Hermann Hofbauer. Steam gasification of various feedstocks at a dual fluidised bed gasifier: Impacts of operation conditions and bed materials. *Biomass Conversion and Biorefinery*, 1(1):39–53, 2011. doi:doi:10.1007/s13399-011-0007-1.

Dale M. Snider. An incompressible three-dimensional multiphase particle-in-cell model for dense particle flows. *Journal of Computational Physics*, 170(2):523–549, 2001. doi:10.1006/jcph.2001.6747.

Dale M. Snider. Three fundamental granular flow experiments and CPFD predictions. *Powder Technology*, 176 (1):36–46, 2007. doi:10.1016/j.powtec.2007.01.032.

Dale M. Snider and Sibashis Banerjee. Heterogeneous gas chemistry in the cpfd eulerian–lagrangian numerical scheme (ozone decomposition). *Powder Technology*, 199(1):100–106, 2010. doi:10.1016/j.powtec.2009.04.023.

Rajan K. Thapa and Britt M. Halvorsen. Heat transfer optimization in a fluidized bed biomass gasification reactor. *Heat Transfer XIII: Simulation and Experiments in Heat and Mass Transfer*, 83:169, 2014.

Li-Qun Wang and Zhao-Sheng Chen. Gas generation by co-gasification of biomass and coal in an autothermal fluidized bed gasifier. *Applied Thermal Engineering*, 59(1):278–282, 2013. doi:10.1016/j.applthermaleng.2013.05.042.

Wen C. Yu. Mechanics of fluidization. In *The Chemical Engineering Progress Symposium Series*, volume 6, pages 100–101, 1966.

Peak Load Cutting in District Heating Network

Petri Hietaharju Mika Ruusunen
Control Engineering, University of Oulu, Finland, {petri.hietaharju,mika.ruusunen}@oulu.fi

Abstract

Simulations of different peak load cutting scenarios in district heating of buildings were performed. Decrease in percentages of 30%, 50%, and 70% in peak loads was analyzed for the two modelled apartment buildings. Simulation results show that even 70% peak load cuts are possible in individual buildings. However, results also reveal that for some buildings 30% peak load cuts would require compromising with the indoor temperature. Therefore, it is important to take into account the different heat storing capacities available in each of the buildings. In future, systems with multiple buildings will be studied to effectively utilize individual heat storing capacities to cut city level peak loads. Simulations presented in this article show that better energy efficiency in district heating can be achieved by predicting the energy consumption and utilizing the thermal mass of a building.

Keywords: district heating, peak load cutting, optimization, indoor temperature prediction, modelling

1 Introduction

Heat represents more than half of the world's total energy consumption and three-quarters of the fuels used to meet this heat demand consist of fossil fuels (Eisentraut and Brown, 2014). Despite these facts, heat is largely ignored in the climate change debate. Nevertheless, it is important to implement new energy efficiency measures in the heat sector. Cutting peak loads in district heating network is one of such measures and in this work peak load cutting is studied with simulations.

Peak loads in district heating network occur when the heat demand exceeds the production capacity of available heating power. This means that reserve power plants need to be started to satisfy the heating demand. This raises production costs (and also environmental impact) for the energy producer as more expensive oil is used for fuel instead of wood, peat or coal. Therefore, it is in the interest of energy companies to cut peak loads and reduce the use of oil. Additionally, CO_2 emissions are also reduced. At the same time, more accurate and stable indoor temperature control could be achieved by implementing the optimization routines for energy consumption.

A concept for peak load cutting has been presented in (Hietaharju and Ruusunen, 2015). The concept aims to cut peak loads by utilizing building thermal mass as a short term heat storage. Building thermal mass and its use in peak load cutting has also been discussed in (Braun, 2013; Henze et al., 2007; Sun et al., 2013; Kensby et al., 2015; Hagentoft and Kalagasidis, 2015; Ståhl, 2009). Braun (2013) presented a review on load control utilizing building thermal mass including simulation, laboratory and field studies. It showed that there is significant saving potential for using building thermal mass, but it is affected by many factors, including utility rates, type of equipment, occupancy schedule, building construction, climate conditions, and control strategy. These factors were further studied by Henze et al. (2007) using a sensitivity analysis. Sun et al. (2013) presented a more recent look into peak load cutting. They found that in existing studies more than 30% daily peak load reduction and also significant overall cost savings from 8.5% to 29% had been achieved utilizing building thermal mass. They also found that there exists model based as well as model free solutions. The amount of energy stored in the building thermal mass is difficult to identify and model based solutions are needed, but they mention the difficulties related to complex physical models and their identification.

Kensby et al. (2015) demonstrated that heavy buildings can tolerate relatively large variations in district heating energy while still maintaining desired indoor temperature. The effect of heating power reduction on indoor temperature was also studied by Hagentoft and Kalagasidis (2015). In case of 1 °C change in the control signal, the indoor temperature drop was below 0.2 °C after 24 hours, which shows that the building thermal mass can be potentially utilized for peak load cutting. However, they mention that the results are highly dependent on the thermal characteristics of the building. In that regard, Ståhl (2009) found that thermal effusivity, which is a function of thermal conductivity and heat capacity and represents the materials ability to exchange thermal energy with its surroundings, is the most important parameter when considering the heat storage capacity of a building. Heavy buildings have higher thermal effusivity and therefore offer higher energy storage capacity compared with light buildings with lower thermal effusivity. Also,

the indoor temperature is typically more stable in heavy buildings.

To utilize thermal mass effectively to cut peak loads and optimize heat consumption, one has to be able to predict the future heat demand. This can be achieved by modelling the thermal behavior of a building, namely the indoor temperature. This way also the quality of living for the residents can be ensured by maintaining the indoor temperature at acceptable level despite the cuts in heating power. In this work, an indoor temperature model (Hietaharju et al., unpublished) was applied to simulate peak load cutting in two apartment buildings. Finally, simulation results for different peak load cutting scenarios are presented and discussed.

2 Data

Two apartment buildings located in the city of Jyväskylä in Finland were studied. Building 1 was constructed in 2011 whereas Building 2 was constructed in 1972 and renovated in 1993. Basic information about the buildings is presented in Table 1. Ground plans and elevation drawings were also available for the studied buildings.

Measurement data for both of the buildings was acquired in early January 2015. Data included hourly values for heating power. In addition, outdoor temperature measurements were recorded for the same time period. Hourly indoor temperature measurements for both buildings were also available. Both buildings contained several indoor temperature measurements which were located in living rooms and hallways.

Table 1. Building Information.

	Building 1	**Building 2**
Year of construction	2011	1972
Year of renovation	-	1993
Floors	4	7
Apartments	75	53
Living space (m^2)	3563	3024
Floor space (m^2)	4200	3703
Volume (m^3)	15617	12400

3 Methods

Previously identified and validated (Hietaharju et al., unpublished) indoor temperature model was used for predicting the indoor temperature evolution over time in the buildings. The model structure (Equation 1) is based on Newton's cooling law and includes heat capacity (C) and heat loss coefficient (U) as physical parameters. Inputs for the model are indoor temperature (T_i), outdoor temperature (T_o), and heating power (P) which can include a lag of k hours. Time step (Δt) for the model is one hour. Model output is the hourly indoor temperature along the defined prediction horizon.

$$T_{i,t} = a\left(T_{i,t-1} + \frac{\Delta t}{C}\left(P_{t-1} - U\left(T_{i,t-1} - T_{o,t-1}\right)\right)\right) + b \quad (1)$$

Initial data described in Section 2, including ground plans and elevation drawings, was used to calculate physical model parameters. Some assumptions were made about the construction materials due to insufficient information. After the calculation of the physical model parameters, input data mentioned before was used to estimate additional model parameters a and b. The indoor temperature model was then utilized to optimize heating power in pilot buildings considering different peak load cutting scenarios.

The scenarios for cutting the peak loads considered simulations for 30%, 50%, and 70% reduction in the heating power. These cuts were made at the morning hours between 7 and 10 am. Load cuts were calculated from the actual measured district heating power. During the simulations, maximum allowed power was restricted accordingly during the peak load hours. Cost function for peak load cutting minimized the power consumption while keeping the indoor temperature between the control limits. This was achieved with a constraint by increasing the cost function value if the indoor temperature would have exceeded the limits according to the model prediction. Also the increase and decrease in the amount of heating power was restricted to prevent too large hourly power changes.

All modelling and optimization work was programmed and evaluated in MATLAB® software with simulations. MATLAB®'s *simulated annealing* algorithm was utilized in peak load cutting simulations to optimize the usage of heat energy per building with respect to the constraint. In all the succeeding simulations, the prediction horizon of 100 hours was applied with acquired data from the buildings.

4 Results and Discussion

Firstly, both apartment buildings were modelled applying the indoor temperature model (Hietaharju et al., unpublished). Modelling results are presented in Table 2. The model performance was evaluated by calculating mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE).

Table 2. Performance of the Indoor Temperature Model.

	Building 1	**Building 2**
MAE (°C)	0.24	0.38
MAPE (%)	1.07	1.69
RMSE (°C)	0.33	0.44

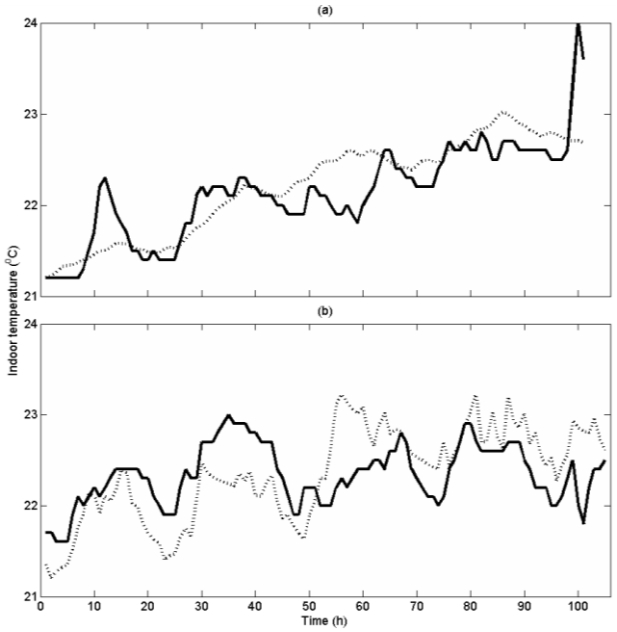


Figure 1. Measured (solid line) and predicted (dotted line) indoor temperature: (a) Building 1 and (b) Building 2.

Figure 1 shows the measured and the modelled indoor temperatures with data from Building 1 and Building 2. The time period for simulations was from January 3rd 2015 to January 7th 2015. For the Building 1, indoor temperature measured from the living room was used as a reference. Respectively, for the Building 2 the temperature measured at the hallway was the model reference. It is important to notice the dynamic behavior that the model manages to capture. Although the modelled indoor temperature somewhat deviates from the measured temperature, the changes and trends are still captured by the model. This model property is further emphasized if the model is to be used for control purposes. Also, MAE, MAPE, and RMSE seem to be reasonably low in case of the two building data sets.

Next, different peak load cutting scenarios were simulated utilizing the identified indoor temperature models. For both of the buildings 30%, 50%, and 70% peak load cuts during morning hours between 7 and 10 am were applied in the simulation. Value of 22 °C was assumed to be the minimum desired indoor temperature for both of the buildings during the simulation. The upper limit for the indoor temperature was set to 23 °C for the Building 2 and to 24 °C for the Building 1. These upper limits were determined from the historical indoor temperature data. Figure 2 and Figure 3 show the simulation results for the peak load cutting. For the indoor temperature, measured and optimized indoor temperatures are presented. The grey bar in the district heating graph represents the peak load period during which the maximum allowed heating power was restricted. Restricted power values were calculated from the measured district heating power by taking the average of the measured district heating power during the peak load period and multiplying this by the desired cut percentage.

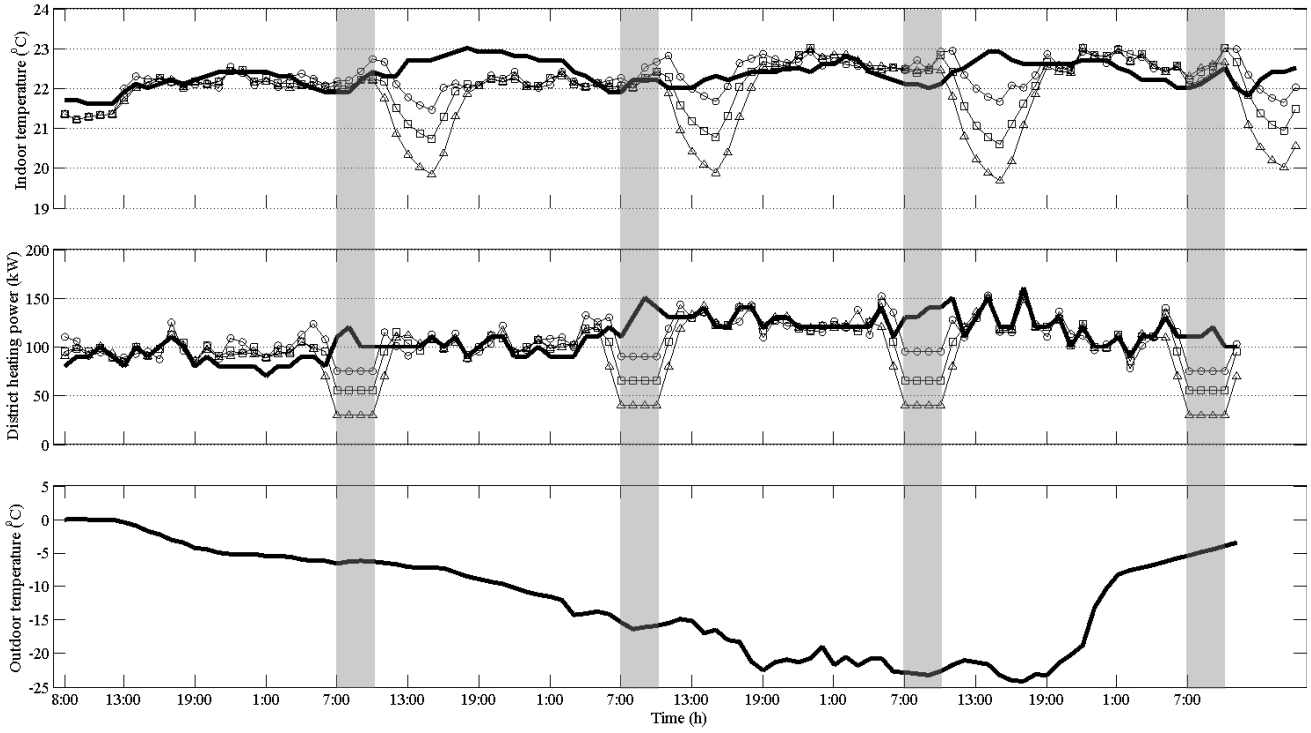


Figure 2. Simulation results for Building 2 in the case of 30%, 50%, and 70% peak load cut during 7-10 am (grey bars). Upper: measured (solid line) and simulated (30%: circle; 50%: square; 70%: triangle) indoor temperature. Middle: measured (solid line) and simulated (30%: circle; 50%: square; 70%: triangle) district heating power. Lower: measured outdoor temperature.

In Figure 2 for the Building 2, a clear change in indoor temperature can be seen as a result of peak load cutting. Delay in the indoor temperature response is due to the lag in the applied model. For the 30% peak load cut it can be seen that at first, the temperature rises to the set upper limit as heat is stored into the building before the peak load period. Then, the temperature drops as the heating power is limited to lower level during the peak loads. In the case of Building 2, the indoor temperature drops below the minimum of 22 °C in all of the simulated peak load cutting scenarios. For 50% and 70% peak load cuts the storage capacity of the building is not enough and the indoor temperature decreases significantly due to the peak load cutting. For the simulation period, the total energy savings are 2.2%, 7.5%, and 15.8% for the 30%, 50%, and 70% peak load cuts respectively.

As can be seen in Figure 3, the simulated change in indoor temperature in the Building 1 is small when compared with simulation results in case of the Building 2 staying between 22 and 23 °C. Nevertheless, the increase in temperature can be observed before the peak load period. Also, it can be clearly seen that the indoor temperature is higher in general on larger peak cuts to be able to perform the peak cut without compromising with the indoor temperature. Unlike in Building 2, the indoor temperature remains above 22 °C in Building 1. Total energy savings for the 50% and 30% peak load cuts for the simulation period were 1.0% and 2.9% respectively. With 70% peak cut energy is not saved during the simulation period but 2.3% more is needed. These results are much lower in comparison with the savings achieved for Building 2. This is partly caused by the high peak in the first four hours in the heating power, which can be seen in Figure 3. This is due to the fact that the initial measured indoor temperature in Building 1 is below the desired level and it has to be raised. If the heating power for these first four hours is excluded from the total energy saving calculations, total energy savings are 6.2%, 4.1%, and 0.7% for the 30%, 50%, and 70% peak load cuts respectively. This shows that energy is saved with every peak load cut percentage, but the savings are still lower, except for 30% cut, when compared with Building 2. This results from the fact that in Building 2 heat is not significantly stored with 50% and 70% peak load cuts as the storage capacity is not enough to compensate for such large cuts and therefore the total energy savings are significantly higher in comparison with Building 1 but the indoor temperature has to be compromised. Higher savings for the 30% peak load cut in Building 1 compared with Building 2 can be explained with larger heat storage capacity in Building 1. In Building 2, 30% peak load cut requires all heat storage capacity to be utilized as can be seen in Figure 2 where the indoor temperature rises to the maximum limit of 23 °C before the peak load cut periods. In Building 1, heating power can be kept on a

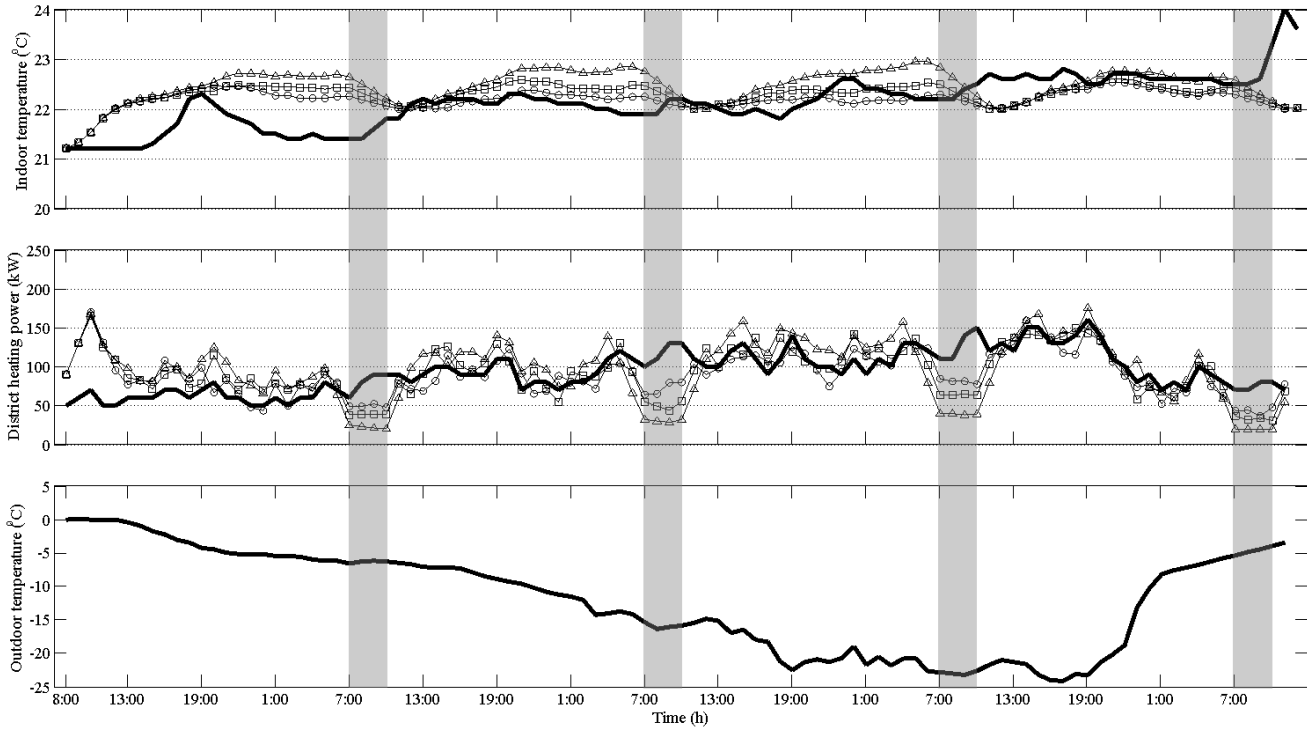


Figure 3. Simulation results for Building 1 in the case of 30%, 50%, and 70% peak load cut during 7-10 am (grey bars). Upper: measured (solid line) and simulated (30%: circle; 50%: square; 70%: triangle) indoor temperature. Middle: measured (solid line) and simulated (30%: circle; 50%: square; 70%: triangle) district heating power. Lower: measured outdoor temperature.

lower level as the heat storage capacity of the building is more than enough to manage 30% cut in peak loads.

It can be seen from the results that the two buildings have different heat storing capacities. In this view, the Building 2 lacks the heat storing capacity to maintain the desired indoor temperature during the peak load cut of 30% or more, which can be clearly seen in Figure 2. However, the indoor temperature could be allowed to drop below the desired level for the duration of the peak loads. In the case of 30% peak load cut, the indoor temperature in Building 2 decreases at most about 0.5 °C below the desired temperature, which could be very well allowed. With 50% peak load cut the indoor temperature drops 1 °C below the desired temperature, which could still be allowed. If temperature drops of the same kind were also allowed for Building 1, the total energy savings would be more significant. On the other hand, also the maximum indoor temperature could be allowed to rise higher to get more heat storage capacity and ensure that indoor temperature does not drop below the desired level, but this could raise the total energy consumption. This shows how important it is to properly define the constraints for the optimization. Also, it is worth noticing that the studied buildings are apartment buildings where the indoor temperature must be more strictly maintained at acceptable level than for example in a school or an office building where the indoor temperature can be allowed to fluctuate more freely during the off-hours. In these kind of buildings, the use of building thermal mass could be even more effective. This is further supported by the preliminary results from an online test performed by the authors in a school building (Hietaharju and Ruusunen, 2015). There the district heating power was optimized for a 24-hour period and it resulted in 14% energy savings for the period and an average of 25% peak load cut during the morning hours. Nevertheless, buildings exist with different heat storing capacities and therefore it is important to investigate multiple buildings as the city level peak loads are desired to be cut. In that case, the peak cutting would be distributed between the buildings and their storage capacities could be effectively used without extensively variating the indoor temperature of individual buildings. Systems with groups of buildings have already been investigated but will be further studied in the future.

5 Conclusions

An indoor temperature model was applied to simulate different peak load cutting scenarios of district heating. According to the simulations, peak cutting potential in the tested two buildings varied because of different heat storing capacities. Nevertheless, in both of the buildings peak loads could be cut 30%, 50%, and even 70%. In the Building 1, performing the peak cutting did not have a significant effect on the indoor temperature and it stayed between the desired levels. In the Building 2, the heat storing capacity was not enough to cut the peak loads by 30% or more without the indoor temperature fluctuating out of the defined limits. This shows the importance of properly defined constraints for the indoor temperature. Furthermore, these simulations demonstrate that the buildings have different heat storing capacities and therefore it is important to investigate systems with multiple buildings. In that case, the load cutting will be distributed between the buildings and their specific heat storing capacities can be effectively used to take into account variation in the temperature dynamics.

All the results presented in this article are simulations based on actual measured data. The next phase would be to conduct field-tests to evaluate the method's performance in a real building environment. In this article, peak load cutting and optimization is considered in a single building, but the overall goal is to develop optimization and peak load cutting methods for a system of multiple buildings. Having multiple buildings changes the picture completely and one can make conclusions about the energy savings also on city level. Systems with multiple buildings have already been investigated by the authors, but will be further studied in the future. Simulations presented in this article show that better energy efficiency in district heating can be achieved by predicting the energy consumption and utilizing the thermal mass of a building.

Acknowledgement

This work has been funded by the Finnish Funding Agency for Innovation (TEKES) as part of the KLEI project (40267/13). Authors thank Jyväskylän Energia Oy and Jyväskylän Vuokra-asunnot Oy for providing data for this study.

References

J. E. Braun. Load control using building thermal mass. *Journal of Solar Energy Engineering*, 125(3):292–301, 2003. doi:10.1115/1.1592184.

A. Eisentraut and A. Brown. *Heating without Global Warming – Market Developments and Policy Considerations for Renewable Heat*. IEA, Paris, France, 2014.

C.-E. Hagentoft and A. S. Kalagasidis. Effect Smart Solutions for District Heating Networks Based on Energy Storage in Buildings. Impact on Indoor Temperatures. *Energy Procedia*, 78:2244–2249, 2015. doi: 10.1016/j.egypro.2015.11.346.

G. P. Henze, T. H. Le, A. R. Florita, and C. Felsmann. Sensitivity Analysis of Optimal Building Thermal Mass Control. *Journal of Solar Energy Engineering*, 129(4):473–485, 2006. doi: 10.1115/1.2770755.

P. Hietaharju and M. Ruusunen. A concept for cutting peak loads in district heating. In *Proceedings of the Automaatio XXI seminar*, Helsinki, Finland, 2015.

P. Hietaharju, M. Ruusunen, and K. Leiviskä. A parametric physical model for indoor temperature prediction and control in buildings. Unpublished.

J. Kensby, A. Trüschel, and J.-O. Dalenbäck. Potential of residential buildings as thermal energy storage in district heating systems – Results from a pilot test. *Applied Energy*, 137:773–781, 2015. doi: 10.1016/j.apenergy.2014.07.026.

F. Ståhl. *Influence of thermal mass on the heating and cooling demands of a building unit*. PhD Thesis, Chalmers University of Technology, Sweden, 2009.

Y. Sun, S. Wang, F. Xiao, and D. Gao. Peak load shifting control using different cold thermal energy storage facilities in commercial buildings: A review. *Energy Conversion and Management*, 71:101–114, 2013. doi: 10.1016/j.enconman.2013.03.026.

Screening of Kinetic Rate Equations for Gasification Simulation Models

Kjell-Arne Solli, Rajan Kumar Thapa, Britt Margrethe Emilie Moldestad

Department of Process, Energy and Environmental Technology, University College of Southeast Norway, Norway
kjell-arne.solli@usn.no

Abstract

The energy from biomass can be utilized through the thermochemical conversion process of pyrolysis and gasification. The process involves solid phase and fluid phase interactions. Computational Particle Fluid Dynamics (CPFD) tools are most commonly used for simulations. The chemical processes involved is described by reaction rate expressions and equilibrium constants. These expressions are often not well studied, but rather adapted from previous studies in lack of better knowledge. Methodology and tools are presented to aim in the selection and optimization of rate expressions for a particular process. Simulation tools for reactions in batch or plug-flow conditions are shown applicable to study selected chemical reactions in detail. Results from one such study is compared to CPFD as well as CSTR results of a gasification process. The reaction scheme for the simulation model could be simplified.

Keywords: gasification, reaction kinetics, KINSIM, Barracuda, Aspen Plus, fluidized bed

1 Introduction

Biomass is the oldest source of energy known to men and contributes to 14% of world's energy consumption (Bain *et al.*, 1998; Saxena *et al.*, 2009). Energy recovery from biomass is possible through the thermochemical conversion processes pyrolysis and gasification. In the gasification process, the biomass undergoes endothermic reactions to produce a mixture of combustible gases that is called producer gas, which can be used in gas engines, gas turbines or fuel cells to produce combined heat and electricity or it can be used as raw material in production of liquid bio fuel.

Different types of reactors can be used in the gasification process. The most efficient types are the fluidized bed reactors, and the efficiency of these reactors mainly depends on the thermo-chemical and fluid dynamic behavior in the reactor. (Olofsson *et al.*, 2005; Pfeifer *et al.*, 2011). The fluid dynamics depends on the gas-solid flow inside the reactor while the thermochemical behavior or reactions and their kinetics depend on the heat supply, the residence time and the gas-particle mixing in the reactor. The efficiency of the biomass-based energy technology has to be further increased to make the technology more sustainable in the world energy market, which is still dominated by fossil fuels. Study of chemical reactions and their kinetics in an operating plant or a reactor model, is difficult due to high operating temperature. Computational models are therefore used to optimize the reactors. In order to have a simulation model that predicts chemical composition, it is crucial to implement reaction schemes with reaction rates and the targeting equilibrium composition that correspond to real life conditions.

Different computational models have been used to simulate the biomass gasification process. Thapa et al. (Thapa and Halvorsen, 2014; Thapa *et al.*, 2014) have simulated the gasification process using the Computational Particle Fluid Dynamics (CPFD) tool, by Barracuda VR15 software. Thapa implemented the set of kinetic equations developed by Snider et al (Snider *et al.*, 2011), Kauhsal et al. (Kaushal *et al.*, 2007), Gomez-Barea et al. (Gómez-Barea and Leckner, 2010) and Umeki et al. (Umeki *et al.*, 2010) together with the stoichiometric equations in Barracuda. The results of the simulations were compared to the measured values from an existing biomass gasification plant. The CPFD methodology solves the fluid and particle equations in three dimensions. The fluid dynamics is described by averaged Navier-Stokes equations with strong coupling to the particle phase. To apply CPFD methodology for heat transfer and chemistry, an enthalpy equation is solved to calculate flows with large chemistry-induced temperature variations. The CPFD method is a hybrid numerical method, where the fluid phase is solved using Eulerian computational grid and the solids are modeled using Lagrangian computational particles (Andrews and O'Rourke, 1996; Snider, 2001).

Eikeland et al. (Eikeland *et al.*, 2015) used the process simulation tool Aspen Plus to simulate the biomass gasification process. The reactions are simulated by using Gibbs reactor or continuous stirred tank reactor (CSTR). The Gibbs reactor only includes the stoichiometric equations, whereas the CSTR model includes both the stoichiometric equations and the related reaction rates for analysis of the producer gas composition and its heating value. Eikeland used the the set of kinetic equations developed by Umeki et al. (Umeki *et al.*, 2010). Aspen Plus is a one-dimensional steady state simulation model.

The aim of the current work is to investigate the sets of reaction scheme kinetics published in literature by simulating them using the ReactLab™ KINSIM software from Jplus Consulting (Norman *et al.*, 2016), and compare to the results of the previous works performed with CPFD and Aspen Plus (Eikeland *et al.*, 2015; Thapa and Halvorsen, 2014).

2 Reaction Kinetics

For a reaction of species A and B to C and D, the rate *r* of the forward reaction is in general given by the equation

$$r = k[A]^a[B]^b \quad (1)$$

where *a* and *b* equals the stoichiometric coefficients for an elementary reaction, while this need not be the case for a global reaction consisting of several elemental steps. For a global reaction, the rate-limiting step may include species not present in the global rate formulation. The brackets denotes the concentration (or activities, to be precise) of species *A* and *B* in specified units. The rate constant *k* is a function of temperature as well as specific conditions like the presence of a catalyst. The temperature dependence of *k* can be described by the modified Arrhenius equation

$$k = BT^n exp^{\frac{-E_a}{RT}} \quad (2)$$

where B is a temperature-independent constant (McNaught and Wilkinson, 1997) and n is a number between –1 and +1. The constant n=0 gives the original Arrhenius equation and then usually the letter A denotes the pre-exponential factor. The constant E_a is the activation energy and *R* is the universal gas constant. The reverse reaction rate is given by a similar rate equation. The thermodynamic equilibrium of a chemical reaction corresponds to forward and reverse rates being equal. Thus, denoting the reverse reaction constants with an apostrophe, the equilibrium constant K is given by

$$K = \frac{k(\text{forward})}{k'(\text{reverse})} = \frac{A}{A'} exp^{\frac{-E_a+E'_a}{RT}} = \frac{[C]^c[D]^d}{[A]^a[B]^b} \quad (3)$$

Some reactions, like proton transfer in aqueous solutions, achieve equilibrium close to instantaneous such that product concentrations can be calculated from the equilibrium constants and mass balance. For most reactions, an equilibrium condition is not reached within available time frame, thus product concentrations must be deduced from differential equations. The solution of these may be by algebraic or stochastic methods.

3 Previous Work

Many of the published reaction rate expressions were developed for simulation of coal gasification processes. They have later been applied to biomass sources. In addition, the study of combustion processes involve many relevant species and expressions, and are thus applied for gasification.

3.1 The Gasification Process

The process for gasification of biomass can be divided in three consecutive steps; namely evaporation, pyrolysis, and gasification of resulting char product and volatiles. Both evaporation and pyrolysis proceeds close to instantaneous at the typical process temperatures of above 700 °C. The products after pyrolysis can be characterized as ash, tar, char and gases. The ash fraction is mostly minerals. The tar fraction is higher molecular weight hydrocarbons including small fractions of other elements. Both the ash and tar fractions are of minor importance for the total simulation, and therefore often skipped in the simulation model. The char fraction is mostly solid carbon (C), and is usually treated as pure carbon in following reactions with the pyrolysis gases. The main gases produced from pyrolysis are water (H_2O), carbon monoxide (CO), carbon dioxide (CO_2), hydrogen (H_2) and methane (CH_4). In addition comes the water produced from evaporation. For a fluidized process to work, a fluidization gas is needed. Steam (H_2O) is used in the process studied in this work, while carbon dioxide or air (oxygen, nitrogen) are commonly used in other processes.

3.2 Kinetic Models

The overall chemical reactions to be included in a simulation model then involve reactions between C, CO, CO_2, H_2, CH_4 and H_2O. Some possible chemical reactions are listed in Table 1 and Table 2. Since oxygen is not added to the system, combustion reactions with oxygen has been omitted.

The table field Referring source points to multiple references. It has become common practice to refer to an author who in turn refers to another author and so on, and it is therefore often difficult to evaluate the relevance of the original source, see discussion below.

3.3 Complexity of Sources for Kinetic Rate Equations

For the rates involving solid phase, reference to the Japanese text of Watanabe (Watanabe *et al.*, 2002) is made by Umeki et al. (Umeki *et al.*, 2010). Snider (Snider *et al.*, 2011) is referenced in recent work by Xie (Xie *et al.*, 2012) and Eikeland (Eikeland *et al.*, 2015). Snider in turn refers to Syamlal and Bissett (Syamlal and Bissett, 1992) who refers to a user's manual of a computer program by Wen et al. (Wen *et al.*, 1982).

Wen et al. are listing 4 sets of parameters for 4 types of coal – the parameters are determined from experimental data by Elgin (Elgin, 1974). Any reference to which type of coal that is relevant has been lost in later references.

Table 1. Chemical Reactions Involving Solid Phase (Carbon).

Reactions ***Heat of reaction at 850 °C, ΔH_R (kJ/mol) (Zanzi et al., 2002)***	**Reaction rate forward (top) and reverse (bottom)[a] (mol/(m³•s), J, K)**	**Referring source**
Steam gasification (water-gas reaction): $C(s) + H_2O \rightleftarrows CO + H_2$ *+118,5*	$= 1{,}272 m_s T exp^{\frac{-22645}{T}}[H_2O]$ $= 1{,}044 \bullet 10^{-4} m_s T^2 exp^{\left(\frac{-6319}{T} - 17{,}29\right)}[H_2][CO]$	Snider – Syamlal – Wen (Snider *et al.*, 2011; Syamlal and Bissett, 1992; Wen *et al.*, 1982)
	$= 2{,}07 \bullet 10^7 exp^{\frac{-220000}{RT}} p_{H_2O}^{0{,}73}$ *units not given in source*	Umeki – Watanabe (Umeki *et al.*, 2010; Watanabe *et al.*, 2002)
CO_2 gasification (Boudouard reaction): $C(s) + CO_2 \rightleftarrows 2\ CO$ *+159,5*	$= 1{,}272 m_s T exp^{\frac{-22645}{T}}[CO_2]$ $= 1{,}044 \bullet 10^{-4} m_s T^2 exp^{\left(\frac{-2363}{T} - 20{,}92\right)}[CO]^2$	Snider – Syamlal – Wen
	$= 1{,}12 \bullet 10^8 exp^{\frac{-245000}{RT}} p_{CO_2}^{0{,}31}$ *units not given in source*	Umeki – Watanabe
Methanation reaction: $0{,}5\ C(s) + H_2 \rightleftarrows 0{,}5\ CH_4$ *-87,5*	$= 1{,}368 \bullet 10^{-3} m_s T exp^{\left(\frac{-8078}{T} - 7{,}087\right)}[H_2]$ $= 0{,}151 m_s T^{0{,}5} exp^{\left(\frac{-13578}{T} - 0{,}372\right)}[CH_4]^{0{,}5}$	Snider – Syamlal – Wen

[a] $m_s = M_{wC} \times [C(s)]$ equals mass of carbon, the approximate char component
M_{wC} = molecular weight for carbon and $[C(s)]$ = molar concentration of solid carbon

Table 2. Chemical Reactions in Fluid Phase.

Reactions ***Heat of reaction at 850 °C, ΔH_R (kJ/mol) (Zanzi et al., 2002)***	**Reaction rate forward (top) and reverse (bottom) (mol/(m³•s) , J, K)**	**Referring source**
Water-gas shift reaction: $CO + H_2O \rightleftarrows CO_2 + H_2$ *-33,6*	$= 2{,}78 \bullet 10^6 exp^{\frac{-1516}{T}}[CO][H_2O]$	Gómez – Biba (Biba *et al.*, 1978; Franks, 1967; Gómez-Barea and Leckner, 2010; Yoon *et al.*, 1978)
	$= 7{,}68 \bullet 10^{10} exp^{\frac{-36640}{T}}[CO]^{0{,}5}[H_2O]$ $= 6{,}4 \bullet 10^9 exp^{\frac{-39260}{T}}[H_2]^{0{,}5}[CO_2]$	Snider – Bustamante (GRI) – Moe (Bradford, 1933; Bustamante *et al.*, 2004; Bustamante *et al.*, 2005; Smith *et al.*; Snider *et al.*, 2011)
	$= 2{,}75 \bullet 10^6 exp^{\frac{-10065}{T}}[CO][H_2O]$ $= 6{,}71 \bullet 10^7 exp^{\frac{-13688}{T}}[CO_2][H_2]$	Wang – Lindstedt – Kuo (Jones and Lindstedt, 1988; Kuo, 1986; Wang *et al.*, 2012)
	$= 2{,}50 \bullet 10^5 exp^{\frac{-16600}{T}}[CO][H_2O]$ $= 2{,}38 \bullet 10^3 T exp^{\frac{-16600}{T}}[CO_2][H_2]$	Umeki – (Corella/Maki) (Bustamante *et al.*, 2005; Corella and Sanz, 2005; Snider *et al.*, 2011; Umeki *et al.*, 2010)
	$= 10^6 exp^{\frac{-6370}{T}}[CO][H_2O]$ $= 1{,}92 \bullet 10^3 exp^{\frac{860}{T}}[CO_2][H_2]$	Corella and more (Corella and Sanz, 2005; González-Saiz, 1988; Simell *et al.*, 1999; Xu and Froment, 1989)
	$= 2{,}50 \bullet 10^5 exp^{\frac{-16600}{T}}[CO][H_2O]$ $= 9{,}43 \bullet 10^6 exp^{\frac{-49500}{T}}[CO_2][H_2]$	Maki – Modell (Maki and Miura, 1997; Modell and Reid, 1974)
	$= 2{,}78 \bullet 10^6 exp^{\frac{-1510}{T}}[CO][H_2O]$ $= 1{,}05 \bullet 10^8 exp^{\frac{-5468}{T}}[CO_2][H_2]$	de Souza-Santos – Gibson – Parent (de Souza-Santos, 2004; Gibson and Euker, 1975; Parent and Katz, 1948)
Methane steam reforming reaction: $CH_4 + H_2O \rightleftarrows CO + 3\ H_2$ *+225,5*	$= 3{,}0 \bullet 10^5 exp^{\frac{-15042}{T}}[CH_4][H_2O]$	Gómez – Lindstedt (Gómez-Barea and Leckner, 2010; Jones and Lindstedt, 1988; Yoon *et al.*, 1978)

Reactions *Heat of reaction at 850 °C, ΔH_R (kJ/mol) (Zanzi et al., 2002)*	Reaction rate forward (top) and reverse (bottom) (mol/(m³•s) , J, K)	Referring source
	$= 3{,}1005 exp^{\frac{-15000}{T}} [CH_4][H_2O]$ $= 3{,}556 \bullet 10^{-3} T exp^{\frac{-15000}{T}} [CO][H_2]^2$	Umeki – (Corella) (Corella and Sanz, 2005; Fletcher *et al.*, 2000; Gómez-Barea and Leckner, 2010; Jones and Lindstedt, 1988; Liu and Gibbs, 2003; Thérien *et al.*, 1987; Umeki *et al.*, 2010)
	$= 3{,}10 \bullet 10^{-5} exp^{\frac{-15000}{T}} [CH_4][H_2O]$ $= 1{,}17 \bullet 10^{-3} exp^{\frac{-47900}{T}} [CO][H_2]^3$	revised Umeki – (Lindstedt)
	$= 9{,}1 \bullet 10^{7} exp^{\frac{-15800}{T}} [CH_4][H_2O]$ $= 5{,}52 \bullet 10^{-6} exp^{\frac{211200}{T}} [CO][H_2]^3$	Maki (Maki and Miura, 1997; Modell and Reid, 1974)

For the water-gas shift reaction, the several sources for the reaction rate expression shows a wide variety of reaction rates. Eikeland et al. (Eikeland *et al.*, 2015) refer to Umeki et al. (Umeki *et al.*, 2010). Umeki at al. refers to Corella and Sans (Corella and Sanz, 2005), but are using the GRI expressions referred by Snider et al. (Snider *et al.*, 2011) from Bustamante (Bustamante *et al.*, 2004; Bustamante *et al.*, 2005). Xie et al. (Xie *et al.*, 2012) refers to Gómez-Barea and Leckner (Gómez-Barea and Leckner, 2010), who refers to Biba et al. (Biba *et al.*, 1978) referring to a book by Franks (Franks, 1967).

Biba et al. (Biba *et al.*, 1978) points out that values determined by experiments for the frequency factor cited in the literature are strictly dependent on the form of the carbonaceous material being examined, on the specific surface area, and on the corresponding value of activation energy. They have chosen mean values of literature data, and then made a sensitivity analysis for their mathematical model. Any reference to which type of coal that is relevant has been lost in later references.

The history for kinetic parameters used for the methane steam reforming reaction is similar to preceding text. In addition, it seems that a typing error is introduced by Corella and Sans (Corella and Sanz, 2005) and later referenced by Umeki et al. and Eikeland et al. Corella and Sans refers to work by Thérien et al. (Thérien *et al.*, 1987) and Liu and Gibbs (Liu and Gibbs, 2003) from Fletcher et al. (Fletcher *et al.*, 2000) which includes a reference to Jones and Lindstedt (Jones and Lindstedt, 1988). The typing errors have been corrected in the "revised Umeki-(Lindstedt) source" in Table 2.

3.4 Chemical Equilibria

For reactions that are kinetically restricted and far from equilibrium conditions, only the forward reaction rate is significant. This should, of course, be verified by comparing calculated forward and reverse rates at the simulation conditions.

The water-gas shift reaction is relatively fast (although slower than oxygen combustion) and both forward and reverse rates as well as the equilibrium constant is relevant for most simulation conditions. Bustamante et al. (Bustamante *et al.*, 2005) points out that the wide range of kinetic parameters found in literature is due to varying catalytic effect from surfaces present at experimental conditions. Typically, carbon surfaces or deposits and nickel-containing steel (like Hastelloy®) as well as minerals in ash are shown to have catalytic effect on the reaction, resulting in high reaction rates. The expression for the equilibrium constant of the water-gas shift reaction from several sources are shown in Table 3. The last line source, Callaghan (Callaghan, 2006), is a recent thesis covering the details about the water-gas shift reaction, and presumably represents reliable knowledge about the equilibrium constant for the reaction.

4 Example Screening of Kinetic Rate Equations

In order to have a simulation model that predicts chemical composition, both the reaction rate kinetics and the targeting equilibrium composition should correspond to real life conditions. Taken the vast difference among literature data, the choice of kinetic rate expressions is not straight forward.

Figure 1 compares the equilibrium constant value, K_{eq}, for the water-gas shift reaction based on selected literature data from Table 3. A positive value of $\log K_{eq}$ corresponds to the forward reaction being favored. The reaction is a moderately exothermic reversible reaction, therefore with increasing temperature the reaction rate increases but the conversion of reactants to products becomes less favorable. It is evident from Figure 1 that some sources are using an equilibrium constant not consistent with reliable data. On the other hand, if equilibrium conditions are not reached, the simulation model may still perform sufficiently well. That is, if the reaction rate expressions can reproduce the chemical reactions at present conditions.

Table 3. Expressions for Equilibrium Constant

Water-gas shift reaction: $CO + H_2O \rightleftharpoons CO_2 + H_2$	
Equilibrium constant, calculated as $\frac{k_{forward}}{k_{reverse}}$ (top) and as referred from source (bottom) (mol/(m³•s), KJ)	***Referring source***
Used: $K = 2{,}65 \bullet 10^{-2} exp^{\left(\frac{3958}{T}\right)}$ (ref Biba) Source (ref to Yoon): $K = 0{,}029 exp^{\frac{4094}{T}}$	Gómez – Biba (Biba *et al.*, 1978; Franks, 1967; Gómez-Barea and Leckner, 2010; Yoon *et al.*, 1978)
Used: $K = 12{,}0 exp^{\left(\frac{2620}{T}\right)}$ Source: $K_{eq} = \exp^{-4{,}33+\frac{4577{,}8}{T}}$	Snider – Bustamante (GRI) – Moe (Bradford, 1933; Bustamante *et al.*, 2004; Bustamante *et al.*, 2005; Smith *et al.*; Snider *et al.*, 2011)
$K = 4{,}10 \bullet 10^{-2} exp^{\left(\frac{2620}{T}\right)}$ (Fitted data, Ref: Kuo)	Wang – Lindstedt – Kuo (Jones and Lindstedt, 1988; Kuo, 1986; Wang *et al.*, 2012)
Used: $K = 4{,}10 \bullet 10^{-2} \frac{1}{T}$ Shown in forward rate: $K = 2{,}65 \bullet 10^{-2} \left(\frac{3958}{T}\right)$	Umeki – (Corella/Maki) (Bustamante *et al.*, 2005; Corella and Sanz, 2005; Snider *et al.*, 2011; Umeki *et al.*, 2010)
Source: $K = 520 exp^{\frac{-7230}{T}}$ below 1123 °C Source: $K = 0{,}0027 exp^{\frac{-3960}{T}}$ above 1123 °C	Corella and more (Corella and Sanz, 2005; González-Saiz, 1988; Simell *et al.*, 1999; Xu and Froment, 1989)
$K = 2{,}65 \bullet 10^{-2} exp^{\left(\frac{32900}{T}\right)}$	Maki – Modell (Maki and Miura, 1997; Modell and Reid, 1974)
$K = 2{,}65 \bullet 10^{-2} exp^{\left(\frac{3958}{T}\right)}$	de Souza-Santos – Gibson – Parent (de Souza-Santos, 2004; Gibson and Euker, 1975; Parent and Katz, 1948)
$\log(K) = -2{,}418 + 0{,}0003855 \times T + \frac{2180{,}6}{T}$	Callaghan (Callaghan, 2006)

The following discussion on reaction rates takes as an example the feed inlet conditions for a fluidized reactor as described by Eikeland (Eikeland *et al.*, 2015) and is comparable by work of Thapa (Thapa *et al.*, 2014). Figure 2 compares the initial reaction rates based on composition at inlet to the reactor and selected reaction scheme. Again, the vast variation in source data is evident. The rate expressions used by Snider from Bustamante is close to the non-catalyzed reaction rates found experimentally by Bustamante (Bustamante *et al.*, 2005). Higher reaction rates presumably are deduced from catalytic conditions. Therefore, a simulation model should include rate expressions from sources that resembles the conditions of the target reactor and biomass feed.

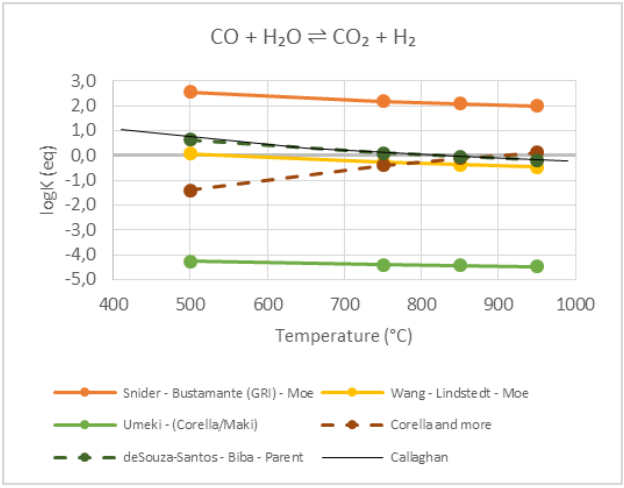


Figure 1. Equilibrium constant for water-gas shift reaction

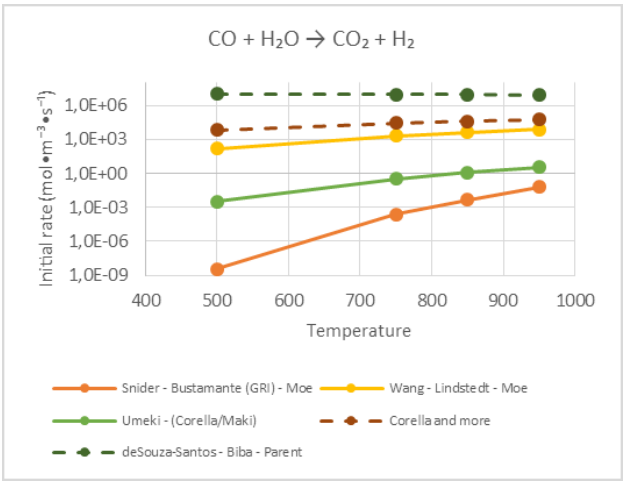


Figure 2. Initial rate of water-gas shift reaction at feed inlet conditions as described by Eikeland (Eikeland *et al.*, 2015).

4.1 KINSIM simulation software

The ReactLab™ KINSIM software from Jplus Consulting (Norman *et al.*, 2016) is one of several software packages available. The software uses MS-Office Excel as frontend to a compiled application of MatLab®. Any defined equilibria is treated as instantaneous, while differential equations for the forward and reverse reaction rates are solved.

4.2 Kinetics

The set of rate expressions used by Thapa (Thapa *et al.*, 2014) together with calculated molar concentrations at inlet feed conditions have been used for time profile simulation of rate expressions. The KINSIM software has been used. The results resemble a batch or plug flow reactor. The residence time 12 s has been chosen to have a conversion comparable to results from Aspen Plus CSTR and Barracuda simulations by Eikeland and Thapa. In addition, simulations with residence time

300 s is shown to get an impression of semi-equilibrium conditions. See Figure 3 to Figure 6.

The Figure 3 to Figure 4 show the water-gas shift reaction to be dominating in consuming CO and water while producing H_2 and CO_2. Still, the semi-equilibrium conditions may not be reached until about 100 s. The total conversion of carbon is strongly limited using this set of rate expressions, and other reactions do not contribute significantly as seen by the very small change in CH_4 concentration. Actually, this imply that a similar simulation result can be obtained by simulating only the water-gas shift reaction. Results including only water-gas shift reaction and water-gas reaction are shown in Figure 5 to Figure 6. At 12 s residence time, the results are identical to the full set of rate expressions. Of course, at 300 s residence time the lacking effect of other reactions are evident, the included carbon conversion reaction is very low.

Figure 7 shows the composition change during the gasification process as simulated at 850 °C by Aspen Plus CSTR, Barracuda CFD and KINSIM plug-flow reactor conditions.

The pyrolysis step corresponds to presumptions for this step made by Thapa and Eikeland (Eikeland *et al.*, 2015; Thapa *et al.*, 2014). In addition, the steam to feed ratio is kept at 1 to 1 weight dry basis. A residence time of 17 s was used for the CSTR conditions, while fluid residence time by Barracuda simulation is estimated to minimum 20 s. The particle residence time is much higher. The last set of results are from using KINSIM with only two rate expressions as shown above. The Barracuda simulation give results that is in between Aspen CSTR and KINSIM plug-flow results while applying the same rate expressions. This is consistent with fluid flow in a fluidized bed reactor be described partly as plug-flow and partly as stirred tank reactor.

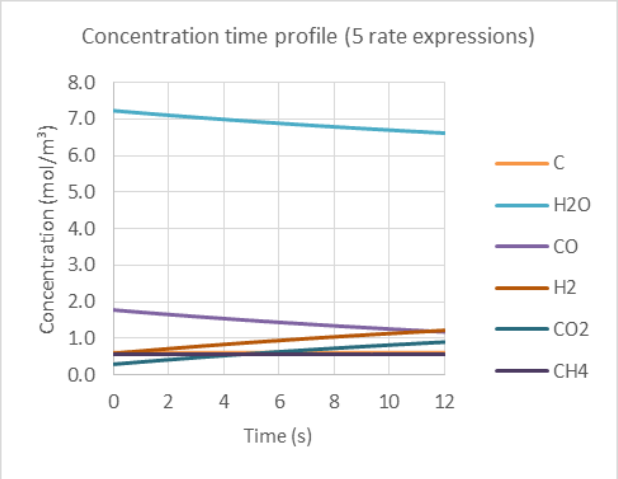


Figure 3. Concentration time profile for the full 5 sets of rate expressions (short time).

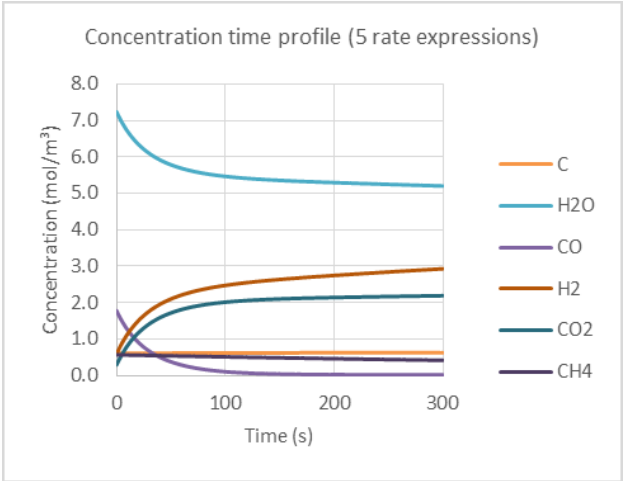


Figure 4. Concentration time profile for the full 5 sets of rate expressions (semi-equilibrium time).

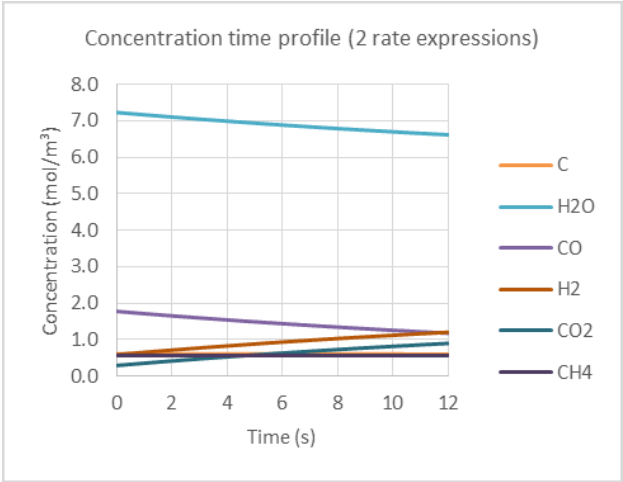


Figure 5. Concentration time profile for only 2 sets of rate expressions (short time).

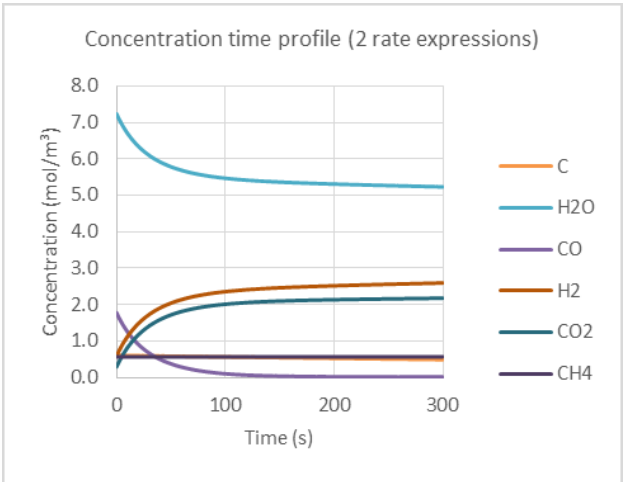


Figure 6. Concentration time profile for only 2 sets of rate expressions (semi-equilibrium time).

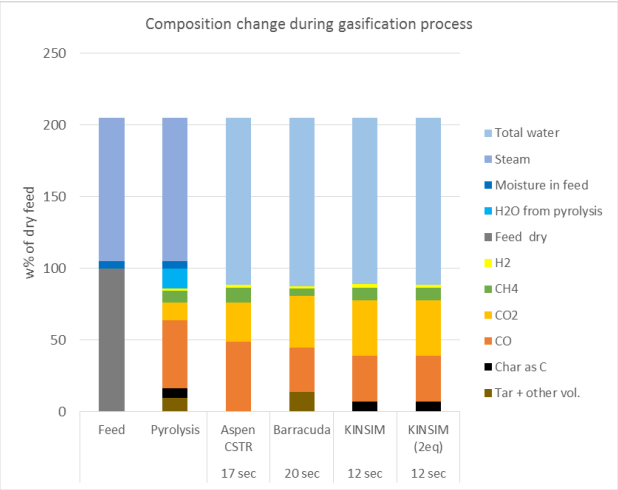


Figure 7. Composition change during a gasification process, feed and steam to outlet stream.

5 Conclusions

The evaluation and screening of candidate rate expressions to be used in a reactor simulation is made difficult by the obscure definition of experimental conditions as background to defined rate expressions. A lightweight tool like KINSIM and others are useful for comparison of several rate expressions and by that evaluate how significant they will contribute to the total simulation. In addition, knowledge about concentrations, residence times and real life results should be used to choose suitable set of rate expressions.

The 5 set reaction scheme studied here could be simplified to 2 sets of reactions by removing those reactions of minimal contribution to overall reaction.

References

M. J. Andrews and P. J. O'Rourke. The multiphase particle-in-cell (MP-PIC) method for dense particle flow. *International Journal of Multiphase Flow*, 22:379-402, 1996.

R. L. Bain, R. P. Overend, and K. R. Craig. Biomass-fired power generation. *Fuel Processing Technology*, 54:1-16, 1998. doi: 10.1016/S0378-3820(97)00058-1

V. Biba, J. Macak, E. Klose, and J. Malecha. Mathematical model for the gasification of coal under pressure. *Industrial & Engineering Chemistry Process Design and Development*, 17(1):92-98, 1978. doi: 10.1021/i260065a017

B. W. Bradford. The water-gas reaction in low-pressure explosions. *J. Chem. Soc.*:1557, 1933.

F. Bustamante, R. M. Enick, A. V. Cugini, R. P. Killmeyer, B. H. Howard, K. S. Rothenberger, M. V. Ciocco, B. D. Morreale, S. Chattopadhyay, and S. Shi. High-temperature kinetics of the homogeneous reverse water–gas shift reaction. *AIChE Journal*, 50(5):1028-1041, 2004. doi: 10.1002/aic.10099

F. Bustamante, R. M. Enick, R. P. Killmeyer, B. H. Howard, K. S. Rothenberger, A. V. Cugini, B. D. Morreale, and M. V. Ciocco. Uncatalyzed and wall-catalyzed forward water–gas shift reaction kinetics. *AIChE Journal*, 51(5):1440-1454, 2005. doi: 10.1002/aic.10396

C. A. Callaghan. *Kinetics and Catalysis of the Water-Gas-Shift Reaction: A Microkinetic and Graph Theoretic Approach*. PhD Thesis: Worcester Polytechnic Institute. Worcester, 2006.

J. Corella and A. Sanz. Modeling circulating fluidized bed biomass gasifiers. A pseudo-rigorous model for stationary state. *Fuel Processing Technology*, 86(9):1021-1053, 2005. doi: 10.1016/j.fuproc.2004.11.013

M. L. de Souza-Santos. *Solid Fuels Combustion and Gasification: Modeling, Simulation, and Equipment Operation*. New York, NY, USA, Marcel Dekker Inc. 2004.

M. S. Eikeland, R. K. Thapa, and B. M. Halvorsen. Aspen Plus Simulation of Biomass Gasification with known Reaction Kinetic. In *Proceedings of the 56th Conference on Simulation and Modelling (SIMS 56).* Linköping, Sweden, Linköping University Electronic Press, Linköpings universitet, pages 149-155, 2015. doi: 10.3384/ecp15119149

D. C. Elgin. Results of Trials of American Coals in Lurgi Pressure Gasification Plant of Westfield, Scotland. In *Sixth Synthetic Pipeline Gas Symposium*. Chicago, Illinois, 1974.

D. F. Fletcher, B. S. Haynes, F. C. Christo, and S. D. Joseph. A CFD based combustion model of an entrained flow biomass gasifier. *Applied Mathematical Modelling*, 24(3):165-182, 2000. doi: 10.1016/S0307-904X(99)00025-6

R. G. E. Franks. *Mathematical Modelling in Chemical Engineering*. New York, N.Y., Wiley, Inc. 1967.

M. A. Gibson and C. A. Euker. Mathematical Modeling of Fluidized Bed Coal Gasification. In *AIChE Meeting*. Los Angeles, CA, USA, 1975.

A. Gómez-Barea and B. Leckner. Modeling of biomass gasification in fluidized bed. *Progress in Energy and Combustion Science*, 36(4):444-509, 2010. doi: 10.1016/j.pecs.2009.12.002

J. González-Saiz. *Advances in biomass gasification in fluidized bed*. PhD Thesis: University of Saragossa. Saragossa, 1988.

W. P. Jones and R. P. Lindstedt. Global reaction schemes for hydrocarbon combustion. *Combustion and Flame*, 73(3):233-249, 1988. doi: 10.1016/0010-2180(88)90021-1

P. Kaushal, T. Proll, and H. Hofbauer. Modelling and simulation of biomass fired duel fluidized bed gasifier at Gussing/Austria. In *International Conference on Renewable Energies and Power Quality.* Sevilla, RE&PQJ, pages 300-306, 2007. doi: 10.24084/repqj05.279

K. K. Kuo. *Principles of Combustion*. New York, John Wiley & Sons. 1986.

H. Liu and B. M. Gibbs. Modeling NH3 and HCN emissions from biomass circulating fluidized biomass gasifiers. *Fuel*, 82:1591-1604, 2003.

T. Maki and K. Miura. A Simulation Model for the Pyrolysis of Orimulsion. *Energy & Fuels*, 11(4):819-824, 1997. doi: 10.1021/ef9601834

A. D. McNaught and A. Wilkinson. In *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book").*

Oxford, Blackwell Scientific Publications. 1997. doi: 10.1351/goldbook

M. Modell and R. C. Reid. *Thermodynamics and its applications*. Englewood Cliffs, N.J, Prentice-Hall. 1974.

S. Norman, P. King, and M. Maeder. *ReactLab KINSIM*. Jplus Consulting Pty Ltd. 2016. Available via http://jplusconsulting.com/products/reactlab-kinsim/

I. Olofsson, A. Nordin, and U. Soderlind. *Initial Review and Evaluation of Process Technologies and System Suitable for Cost-Efficient Medium-Scale Gasification for Biomass to Liquid Fuels*. Master Thesis: University of Umeå, 2005.

J. D. Parent and S. Katz. *Equilibrium Compositions and Enthalpy Changes for the Reaction of Carbon, Oxygen and Steam*. IGT Research Bulletin, 1948.

C. Pfeifer, S. Koppatz, and H. Hofbauer. Steam gasification of various feedstocks at a dual fluidised bed gasifier: Impacts of operation conditions and bed materials. *Biomass Conversion and Biorefinery*, 1:39-53, 2011. doi: 10.1007/s13399-011-0007-1

R. C. Saxena, D. K. Adhikari, and H. B. Goyal. Biomass-based energy fuel through biochemical routes: A review. *Renewable and Sustainable Energy Reviews*, 13:167-178, 2009. doi: 10.1016/j.rser.2007.07.011

P. A. Simell, E. K. Hirvensalo, S. T. Smolander, and A. O. Krause. Steam reforming of gasification gas tar over dolomite with benzene as a model compound. *Industrial and Engineering Chemistry Research*, 38:1250, 1999.

G. P. Smith, D. M. Golden, M. Frenklach, N. W. Moriarty, B. Eiteneer, M. Goldenberg, C. T. Bowman, R. K. Hanson, S. Song, W. C. Gardiner, Jr, V. V. Lissianski, and Z. Qin. *GRI-Mech 3.0*. Available via http://www.me.berkeley.edu/gri_mech/

D. M. Snider. An Incompressible Three-Dimensional Multiphase Particle-in-Cell Model for Dense Particle Flows. *Journal of Computational Physics*, 170:523-549, 2001. doi: 10.1006/jcph.2001.6747

D. M. Snider, S. M. Clark, and P. J. O'Rourke. Eulerian–Lagrangian method for three-dimensional thermal reacting flow with application to coal gasifiers. *Chemical Engineering Science*, 66(6):1285-1295, 2011. doi: 10.1016/j.ces.2010.12.042

M. Syamlal and L. A. Bissett. *METC Gasifier Advanced Simulation (MGAS) model*. 1992. Retrieved from: http://www.osti.gov/scitech//servlets/purl/10127635-7FCHVc/

R. K. Thapa and B. M. Halvorsen. Stepwise analysis of reactions and reacting flow in a dual fluidized bed gasification reactor. In *10th International Conference on Advances in Fluid Mechanics.* Spain, WIT Transactions on Engineering Sciences, pages 37-48, 2014. doi: 10.2495/AFM140041

R. K. Thapa, C. Pfeifer, and B. M. Halvorsen. Modeling of reaction kinetics in bubbling fluidized bed biomass gasification reactor. *The International Journal of Energy and Environment*, 5(1):35-44, 2014. Retrieved from: http://www.ijee.ieefoundation.org/

N. Thérien, P. Marchand, A. Chamberland, and G. Gravel. Computer modeling and simulation of a biomass fluidized bed gasifier. In *Proceedings of the XVIII Congress: The Use of Computers in Chemical Engineering-CEF87.* Gianardi Naxos, Italy, pages 187–192, 1987.

K. Umeki, K. Yamamoto, T. Namioka, and K. Yoshikawa. High temperature steam-only gasification of woody biomass. *Applied Energy*, 87(3):791-798, 2010. doi: 10.1016/j.apenergy.2009.09.035

L. Wang, Z. Liu, S. Chen, and C. Zheng. Comparison of Different Global Combustion Mechanisms Under Hot and Diluted Oxidation Conditions. *Combustion Science and Technology*, 184(2):259-276, 2012. doi: 10.1080/00102202.2011.635612

H. Watanabe, M. Ashizawa, M. Otaka, S. Hara, and A. Inumaru. *Development on numerical simulation technology of heavy oil gasifier*. CRIEPI reports W01023, 2002.

C. Y. Wen, H. Chen, and M. Onozaki. *User's manual for computer simulation and design of the moving-bed coal gasifier. Final report*. 1982. Retrieved from: http://www.osti.gov/scitech//servlets/purl/6422177-trF8L6/

J. Xie, W. Zhong, B. Jin, Y. Shao, and H. Liu. Simulation on gasification of forestry residues in fluidized beds by Eulerian–Lagrangian approach. *Bioresource Technology*, 121:36-46, 2012. doi: 10.1016/j.biortech.2012.06.080

J. Xu and G. F. Froment. Methane steam reforming, methanation and water–gas shift: I. Intrinsic kinetics. *AIChE Journal*, 35(1):88-96, 1989.

H. Yoon, J. Wei, and M. M. Denn. A model for moving-bed coal gasification reactors. *AIChE Journal*, 24(5):885-903, 1978. doi: 10.1002/aic.690240515

R. Zanzi, K. Sjöström, and E. Björnbom. Rapid pyrolysis of agricultural residues at high temperature. *Biomass and Bioenergy*, 23(5):357-366, 2002. doi: 10.1016/S0961-9534(02)00061-2

Model Predictive Control for Field Excitation of Synchronous Generators

Thomas Øyvang[1] Bernt Lie[1] Gunne John Hegglid[2]

[1]Faculty of Technology, University College of Southeast, Norway, {Thomas.Oyvang,Bernt.Lie}@usn.no
[2]Skagerak Energi AS, Norway, GunneJohn.Hegglid@skagerakenergi.no

Abstract

This paper describes a Model Predictive Control (MPC) system for voltage control through field excitation of hydroelectric generating units. An attractive feature of MPC is its capability to handle Multiple Input, Multiple Output (MIMO) systems and nonlinear systems taking constraints into account. The system under study is a power system based on detailed models from Matlab's SimPowerSystems™ and parametrized according to the Nordic model from the Norwegian Transmission System Operator (TSO), Statnett. The primary role of the field excitation control system is to quickly respond to voltage disturbances occurring in the power system. The control system is tested for both first-swing transient stability and long term voltage stability.

Power system modeling and control, model predictive control, SimPowerSystems, first-swing rotor enhancement, long term voltage stability, fmincon

1 Introduction

In this paper a Model Predictive Controller (MPC) for field excitation of synchronous generators is tested, and compared to a classical controller typically used for this purpose. MPC is an advanced control methodology that has proved to be successful in real-life applications. An attractive feature of MPC is its capability to handle Multiple Input, Multiple Output (MIMO) systems and nonlinear systems taking constraints into account (Maciejowski, 2002).

There are mainly two requirements for successful operation of a power system (Hegglid, 1983). The first requirement of reliable service is to keep the generators running in parallel (synchronous) and with necessary capacity to meet load demand. If at any time the generator loses synchronism with the rest of the system, significant voltage and current fluctuations may occur and transmission lines may be automatically tripped by their relays at undesired locations (Kundur, 1994).

A second requirement of reliable service is to maintain the integrity of the power network. Interruptions in this network may hinder flow of power to the load, leading to severe blackouts of the power system. This usually requires a study of large geographical areas since almost all power stations and load centres are connected in one system.

A controller should maintain both of these requirements: keeping the generators running in parallel and to maintain the integrity of the power system. This is achieved mainly by reducing the first swing of the rotors of the synchronous generators after large disturbances, and the damping of power oscillations (also small disturbances). Another important requirement for a controller is to provide necessary reactive power supply Q for enhancement of voltage stability. Reactive power can be used to compensate for voltage drops, but must be provided closer to the demands than active power P needs due to transportation limitations of reactive power through the grid.

A flexible power factor control on large synchronous generators located close to points of high demand could enhance the voltage stability of a power system. The Norwegian Network Code FIKS [1] states that synchronous generators ≥ 1 MVA must connect to the grid with a $\cos\phi$ ≥ 0.86 overexcited and ≤ 0.95 underexcited at maximum load (independent of the location from the point of demand). However, necessary enhancement of voltage stability could be secured through use of more advanced control where more reactive power is temporary available from large generators generated locally.

The paper is organized as follows. Section 2 provides a brief overview of field excitation control of synchronous generators. Section 3 describes MPC for the excitation system and the modeling workflow used in this paper. Section 4 describes MPC tuning. Section 5 introduces, describes and discuss the different tests and results from simulations. In Section 6, conclusions and future perspectives are presented.

2 Field Excitation Control

2.1 Capability Curve

Synchronous machines are capable of producing and consuming reactive power. When the machine is overexcited, it generates Q and deliver it to the power system. Negative Q then, flows from the system into the machine to maintain its magnetization when its own field is underexcited. Generally, reactive power support is divided into

[1]Funksjonskrav I Kraft-Systemet 2012 from the Norwegian Transmission System Operator (TSO) (Statnett, 2012)

two categories: static and dynamic (Kundur, 1994). Dynamic reactive power is produced from equipment that can quickly change the Q independent of the voltage level such as a synchronous generator or condenser (generator without active power exchange P with the grid). Thus, the equipment can increase its reactive power production level when the voltage drops, and prevent a voltage collapse.

A generators operating constraints can be visualized through a capability diagram (Farnham and Swarthout, 1953). In Figure 1, the capability curve is shown for the synchronous generator used for simulations in this paper.

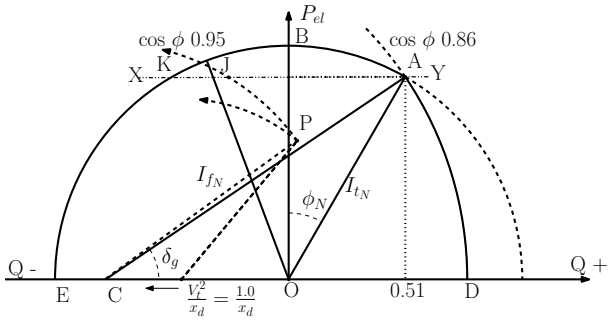


Figure 1. Capability curve for synchronous generator used in simulations. Point A represents the name-plate rated conditions with power factor $\cos\phi = 0.86$ for the generator. Rated (nominal) MVA of the machine $S_N = \sqrt{P_N^2 + Q_N^2}$ is taken as 1.0 per unit on its own rated MVA base. Hence, the rated conditions for machine operation are 0.86 per unit MW and 0.51 per unit MVAr ($\cos\phi = 0.86$, $\sin\phi = 0.51$).

Point A is just one point in a rather extensive area of Figure 1. Few machines are operated at any length of time exactly in the condition of A. The operating conditions in overexcited mode is bounded by the armature current limit AB (circle having its centre at O, and radius equal to rated armature current, I_{t_N} or OA) and the field current limit AD (circle having its centre at C, and radius equal to rated full load field current, I_{f_N} or CA). The controller can control the field current represented by CP in Figure 1. CP/CA is simply the proportion of rated field current, and this is just the amount necessary to permit the machine to handle the P and Q represented by point P. If the load increases without any change in the field current, this causes a movement of operating point P along the arc PK. This path runs almost directly into the instability region (Farnham and Swarthout, 1953) of the generator, e.g minimum field current for stable operation. The final operating point would lie on the XY line, representing limits set by the turbine power. The figure also shows the effect of increased synchronous reactance x_d of the machine. This increase in x_d makes the machine reach the stability limits even faster. The maximum reactive power that can be delivered from this generator is defined by OD=CA-CO (e.g. as an synchronous condenser). With a synchronous reactanse x_d of 1.24 in per-unit, terminal voltage $V_t = 1.0$ in per-unit, and the given operating conditions in Figure 1, the rated field current [2] can be calculated to be 1.6 from

$$I_{f_N} = \sqrt{\left(\frac{V_t^2}{x_d} + \frac{Q_t}{S_N}\right)^2 + \cos\phi_N{}^2}. \quad (1)$$

2.2 Classical Control

The excitation (or field) current required to produce the magnetic field inside the generator, is provided by the exciter and is controlled by an Automatic Voltage Regulator (AVR) (Kundur, 1994). The AVR regulates the generator terminal voltage by controlling the amount of current supplied to the generator field winding by the exciter. Power System Stabilizers (PSS) are feed-forward supplementary control devices which are installed in generator excitation systems to increase damping of (power) oscillations. The specification of excitation systems is guided by IEEE standards 421 (IEEE, 2007). In this paper, a synchronous machine voltage regulator and exciter based on the IEEE type ST1A excitation and Kundur's (Kundur, 1994) generic PSS is used. This type represents a classical exciter model of static potential-source controlled-rectifier systems. This classical control structure is shown in Figure 2

There are two key factors that define an excitation system: the transient gain and the ceiling force ratio (IEEE, 2007). Because of a very high field forcing capability of the system, a field current limiter is employed to protect the generator rotor and exciter. The transient gain has a direct impact on small signal and dynamic stability. Too small a value may fail to give the desired performance, while too high a value (faster response) may result in instability during faults.

3 Concept and Formulation of MPC

MPC is an algorithm where an optimal control problem is solved at the current time, then a receding/sliding horizon technique is applied as time progresses. The predictive controller considers both past situations (given by state) and the changing of the system in a finite future time horizon. To solve the optimal control problem, an optimization routine is needed. In the optimization problem, an objective/criterion to be maximized/minimized is formulated together with constraints (Maciejowski, 2002).

Optimal control is an open loop optimization problem. If input disturbances and references change in the future, the controller has no knowledge about this change since no feedback is presented in the solution. Under such conditions optimal control problems may give good performance. To come around this challenge, *feedback* is introduced to the optimal controller. A way to do this is called Model Predictive Control (MPC) (Sharma, 2014).

[2] The rated field current is the direct current in the field winding of a machine, when operating at rated voltage, current and speed and at rated power factor for synchronous machines (Farnham and Swarthout, 1953).

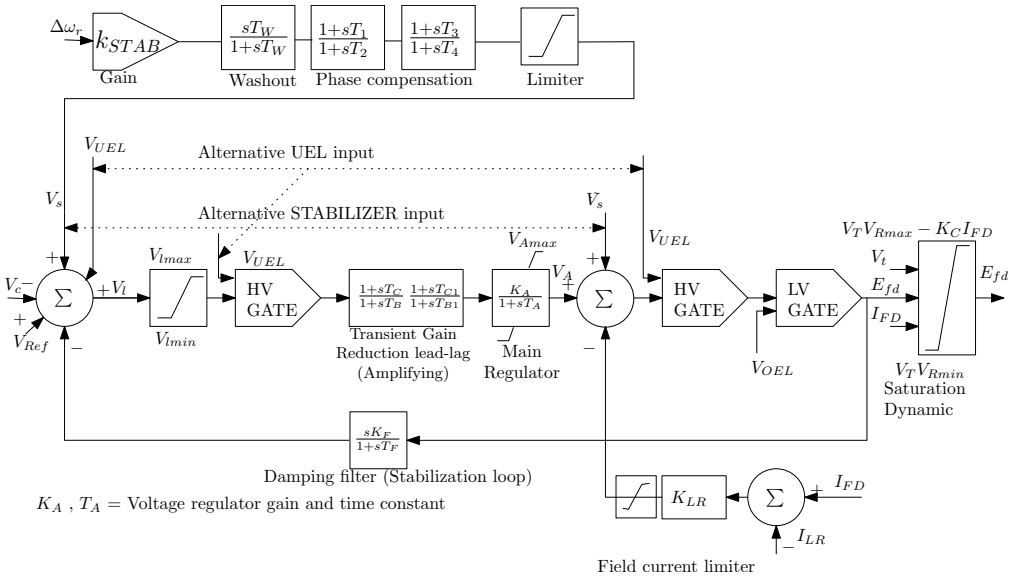


Figure 2. The classical static excitation control ST1A with PSS used in simulations. The PSS representation consist of a phase compensation block, a signal washout block and a gain block. The models are tuned according to Kundur (Kundur, 1994).

As mentioned above, MPC is optimal control with a sliding horizon strategy, e.g. a new optimal control problem is solved at every time step Δt.

Assume the initial state of the plant x_j is known, together with current/future references r_{j+1},r_{j+2},..,r_{j+N} and disturbances w_j, w_{j+1},..,w_{j+N-1}. By solving the optimal control problem, this leads to an optimal open loop control input sequence $u^*_{j|j}$, $u^*_{j+1|j}$,..., $u^*_{j+N-1|j}$. Subscript $k|j$ implies the control input at time k, when current state x_j is known. It is clear that $u^*_{k|j} = u^*_k(x_j, r_{j+1}, .., r_{j+N}, w_j, .., w_{j+N-1})$. The dependence of $u^*_{k|j}$ on $r_{j+1}, .., r_{j+N}, w_j, .., w_{j+N-1}$ gives feed forward from $r_{j+1}, .., r_{j+N}, w_j, .., w_{j+N-1}$. However, optimal control where $u_k = u^*_{k|j}$ does not give feedback since u_k only depends on x_j. To have feedback, we require that u_k depends on x_k. To achieve feedback, introduce receding horizon: set $u_j = u^*_{j|j}$. Then find or estimate x_{j+1} after u_j has been injected, shift/recede the optimal control horizon one step. The process is then repeated to compute $u^*_{j+1|j+1}, u^*_{j+2|j+1}, .., u^*_{j+N+1|j+1}$ by setting $u_{j+1} = u^*_{j+1|j+1}$ and we introduce feedback.

Because of feedforward the controller can react before known disturbances (or set point changes) affect the process. A model of the disturbance should then be included along with the model of the process while solving the optimal control problems at each time step.

3.1 MPC as Excitation Control

For the power system generator excitation control, the selected system outputs are the generator voltage and angular velocity. The manipulating input, u, is the generator field excitation voltage E_{fd} when it is used as a primary controller. Alternatively MPC is also used as a secondary controller to change set-point of the classical (primary) controller $u = V_{ref}$. In Figure 3, the control structure used in the simulations is shown. For transient stability, MPC works as primary control while during long-term stability analysis MPC was tested both for primary and secondary control. The control system can also change between classical and MPC control during simulation.

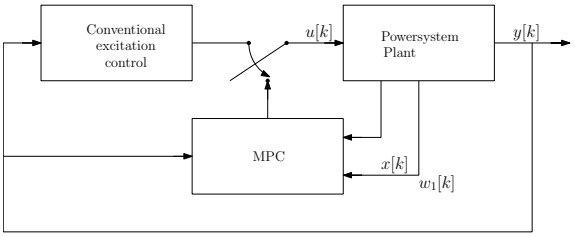


Figure 3. Structure of plant and controller. Here, MPC is a Multiple Input, Single Output (MISO) controller.

The performance index J is scalar and is computed as a summation of the square of the deviations between outputs and references, Δe, and the variations of the controlling inputs to the controlled system, Δu_c. Weighting factors $\mathbf{Q}$, $\mathbf{P}$ and $\mathbf{R}$ are included in J as tuning parameters.

$$J = \sum_{j=1}^{N_p} \Delta e_{V,j}{}^T \mathbf{Q} \Delta e_{V,j} + \Delta e_{\omega,j}{}^T \mathbf{P} \Delta e_{\omega,j} + \sum_{j=1}^{N_c} \Delta u_{c,j-1}{}^T \mathbf{R} \Delta u_{c,j-1} \quad (2)$$

where

$$\Delta u_{c,j} = u_{c,j} - u_{c,j-1} \quad (3)$$

$$\Delta e_{\omega,j} = \omega_{ref} - \omega_{,j} \quad (4)$$

$$\Delta e_{V,j} = V_{ref} - V_{t,j} \quad (5)$$

The control input is the field excitation voltage

$$u_{c_j} = [E_{fd_{c_j}}] \quad (6)$$

N_p and N_c are prediction and control horizon lengths. The inequality constraint on the controlled variable are $E_{fd_c} \in [E_{fd_{min}}, E_{fd_{max}}]$ where $E_{fd_{min}} = -3.2$ per-unit and $E_{fd_{max}} = +3.2$ per-unit (Two times the rated field current according to FIKS (Statnett, 2012)).

The control output

$$y_j = [V_{t,j}, \omega_{g,j}]^T \tag{7}$$

3.2 Modeling and Control Workflow

The modeling part in this paper is done with the *powerlib* library of *SimPowerSystems*™ built on the *Simscape*™ language, running within the *Simulink*® environment. *SimPowerSystems* provides component libraries and analysis tools for modeling and simulating electrical power systems. Since Simulink uses MATLAB as its computational engine, MATLAB toolboxes are available. In this paper, MPC is implemented using the MATLAB Optimization Toolbox and the *fmincon* solver with the *Active-Set* method.

For simulations, the phasor domain solution method with Simulink®variable-step solvers (ode23t and ode23tb) are used. The phasor solution method is mainly used to study electromechanical oscillations of power systems, which is the case study in this paper.

To further increase the speed of optimization, all simulations were done in *Accelerator* mode and with the *Fast Restart* command. The Accelerator mode generates and links code into a C-MEX S-function. The idea of *Fast Restart* is to perform the model compilation once and reuse the compiled information for subsequent simulations. Also the *SimState* function was used for easily integration within the optimization algorithm. A *SimState* is the snapshot of the state of a model at a specific time during simulation.

4 Tuning the MPC Controller

4.1 Time Response

The voltage controller needs to operate freely and without unnecessary restriction within performance limits of the generator and excitation system. A standard procedure for evaluating the response of the closed-loop excitation control system is to document its dynamic characteristics. A small-signal performance measure is expressed in terms of indices associated with time response. A step-response test is done on the regulator in open circuit conditions according to TSO. The mathematical model design to be used for the MPC tuning and also first-swing tests is a classical SMIB (Singel Machine Infinite Bus). This model consist of a three-phase salient-pole synchronous machine modelled in the dq rotor reference frame, three-phase transformer, transmission line and a voltage source as an inifinite bus as shown in Figure 4.

The model is a power plant in Norway: the 175 MVA hydro power plant Kobbelv aggregate 2 owned by Statkraft AS (G2 in Area 2 in Figure 11). The main data for this generator is given in Table 1

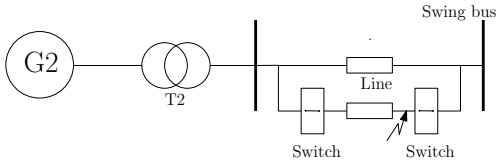


Figure 4. Singel Machine Infinite Bus (SMIB) for transient stability simulations.

Table 1. Main data for G2 in area 2.

Description	Parameter	Value	Unit
Rated power	S_N	175	MVA
Rated Power factor	$\cos\phi_N$	0.86	
Rated voltage	Vt_N	16.5	kV
Frequency	f	50	Hz
Number of polepairs	p	8	
Inertia constant	H	2.9	s

4.2 Open Circuit Conditions

The voltage regulator should be verified by the impulse response in open circuit conditions. For the voltage regulator, a 5% up and down step-response test on the voltage regulator should be carried out. The voltage should be measured over the generator terminals and the response should be non-oscillatory with a overshoot less than 15% of the impulse response itself. The results are shown in Figure 5.

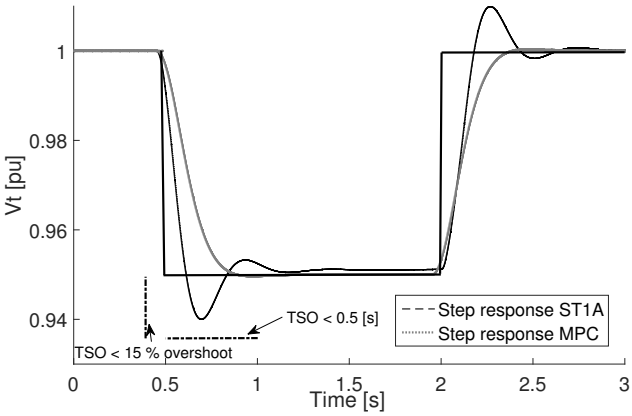


Figure 5. The figure is showing a open circuit small-signal time response test with respect to TSO requirement for both MPC and classical control ST1A. The settling time must be < 0.5 seconds for a 0.95 to 1 step-up and a 1 to 0.95 step-down from steady state value according to FIKS (Statnett, 2012). The overshoot shall be less than 15 % of the change. The MPC tuning parameters were $N_p = 5$, $N_c = 1$, $\mathbf{Q} = 300$, $\mathbf{P} = 5$, $\mathbf{R} = 0.01$ and $\Delta t = 0.05$.

5 First-swing Angle Stability Enhancement

First-swing angle stability (or transient stability) enhancement is necessary to avoid loss of synchronism. Two factors which indicate the stability are the angular swing (during and after a fault) and the critical clearing time t_{crit} [3] (or clearing angle δ_{crit}) of a fault (Grainger and Stevenson, 1994). The generator rotor angle swing normally peaks between 0.4 and 0.75 seconds. This short time demands a fast acting voltage regulator to boost the internal voltage through the field excitation. The steady-state power-angle characteristic presented in Figure 6 shows the highly non-linear relationship between interchange electrical power P_e and angular position δ of the rotors of the synchronous generators (The effect of AVR and damping windings (Kundur, 1994) are neglected).

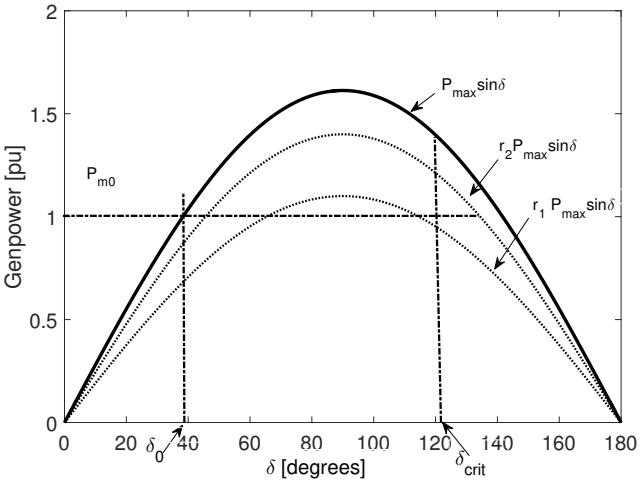


Figure 6. Plot of power-angle curves (Pre-fault P_{max}, during fault $r_1 P_{max}$ and post-fault $r_2 P_{max}$) for synchronous generator G2 Area 2 showing, initial mechanical power and electrical power $P_{m0} = P_e = 1pu$, inital angular position of the rotor $\delta_0 = 39$ degrees and the critical clearing angle calculated to $\delta_{crit} = 122$ degrees and corresponding critical clearing time $t_{crit} = 0.28$ seconds.

Position $\delta_0 < 90$ in Figure 6 is the initial operating point of a stable operation. The swing equation for the machine with constant flux linkage may be written in acceleration form as

$$\frac{H}{180f}\frac{d^2\delta}{dt^2} = P_m - P_e = 1.0 - P_{max} sin\delta \tag{8}$$

$$P_{max} = \frac{E_g \cdot V_{bus}}{x_s} \tag{9}$$

E_g is the synchronous machine transient voltage, x_s is the series transfer reactance between E_g and swing bus V_{bus}, and f is the electric frequency. In all the simulations,

[3] The critical clearing time is the maximum elapsed time from the initiation of the fault until its isolation such that the power system is stable.

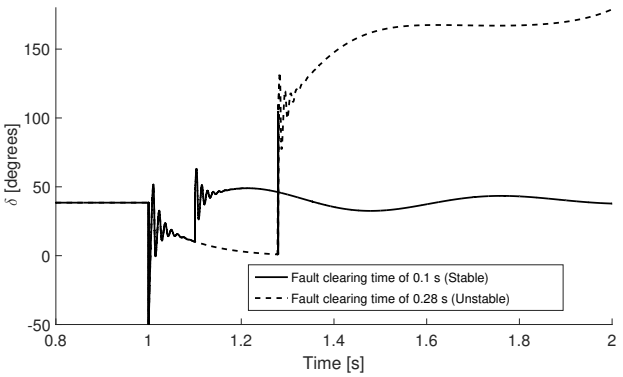


Figure 7. Plot of rotor angle δ affected by a three-phase fault under constant field voltage control for synchronous generator G2 in Area 2. The plot shows two different fault clearing times. The fault cleared after 0.28 seconds lead to loss of synchronism as calculated from t_{crit}.

P_m is kept constant. When P_m equals P_e, the machine operate at steady state synchronous speed. Both the inertia constant H and transient reactance of the machine x'_d has a direct impact on the first swing (transient) studies. A smaller H gives a larger angular swing. P_{max} decreases as the transient reactance increases since it forms a part of the overall series reactance. A decreased P_{max} constrains the machine to swing through a smaller angle from its original position before it reaches the critical clearing angle.

The MPC controller was tested and compared to classical control as shown in Figure 8 and Figure 9.

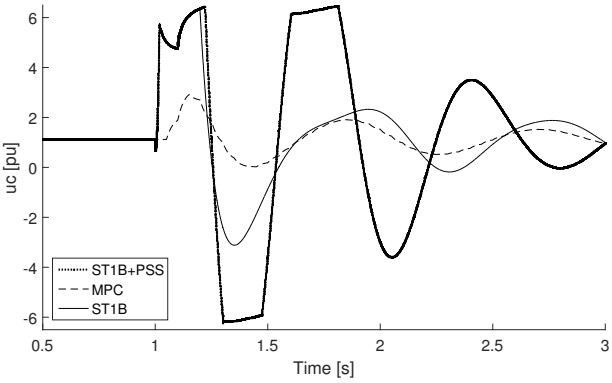


Figure 8. Change in field voltage E_{fd} as a function of time under disturbance of a three-phase arc fault for three different types of control actions. Clearing time was 0.1 second. Future disturbance is not known for the controller.

6 Long-term Voltage Stability Enhancement

6.1 Steady-state Voltage Stability

The steady-state voltage-power characteristics (also called onion surface) shown in Figure 10 for the SMIB, gives insight into the voltage stability problem (Larsson, 2000).

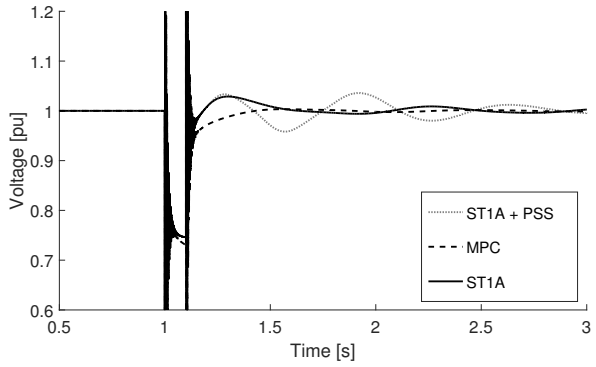


Figure 9. Change in terminal voltage V_t as a function of time under disturbance of a three-phase arc fault for three different types of control actions. Clearing time was 0.1 second. Future disturbance is not known to the controller.

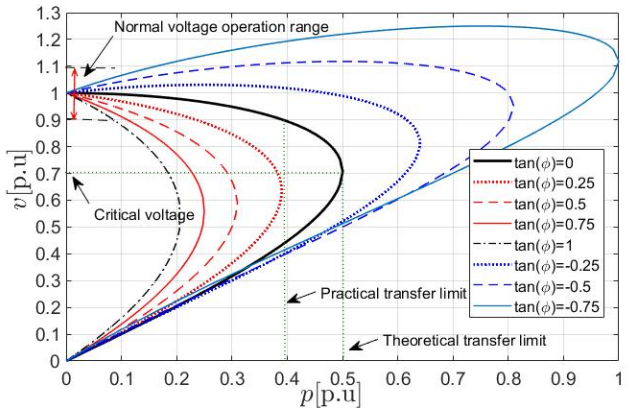


Figure 10. Normalized power(p)-voltage(v) curves (onion surface) for steady-state voltage stability analysis. The practical and theoretical transfer limits and the critical voltage is given for $\tan(\phi) = 0$.

The critical voltage V_c (voltage collapse) points out the theoretically stability limit of the SMIB. The most important factor to provoke a voltage collapse (blackout) is the load model as presented in (yvang et al., 2014) . A widely used dynamic load characteristic is an exponential load model with fractional load exponents combined with time constants for both active and reactive power (Cutsem and Vournas, 1998). If the voltage is lower than a specified value, the load impedance is kept constant. An impedance load model adapts to the voltage. However, a dynamic load that recovers over time combined with automatic tap-changers would stress the power system. When load recovers in a highly loaded system, the need for reactive power increases with I. This will automatically bring the system to the edge. In addition one needs to make sure that generator exciters are limited (Kundur, 1994).

6.2 Power System Simulator

Large-disturbance (e.g. loss of load or loss of generation) long term voltage stability simulations requires the examination of the dynamic performance of the system over a period of time sufficient to capture the interactions of such as tap-changers and field current limiters (Kundur, 1994). This means that it is not only sufficient to simulate a three phase fault with clearing of fault after some milliseconds as done with first swing simulations. After an arc fault the system will either be transiently unstable, partly or as a whole (collapse), or it will return to a stable point. Thus, additionally outage of a line or any reactive power source is of interest (Cutsem and Vournas, 1998). For Long Term Voltage Stability (LTVS) simulations, a more complex simulator was needed to test the MPC controller in contrast to the SMIB simulator in transient studies. A Dynamic Study Model has been developed based on the Nordic Model 2010 High Load Case (Norgesmodellen) from Statnett. The geographical area is in the North region of Norway containing four synchronous generators and one synchronous compensator. It also contains an tap-changer connected to a dynamic load for stressing the system. The model was tuned to operate near its capacity limit and could exhibit cascading failures which could lead to blackouts. The power system model is shown in Figure 11.

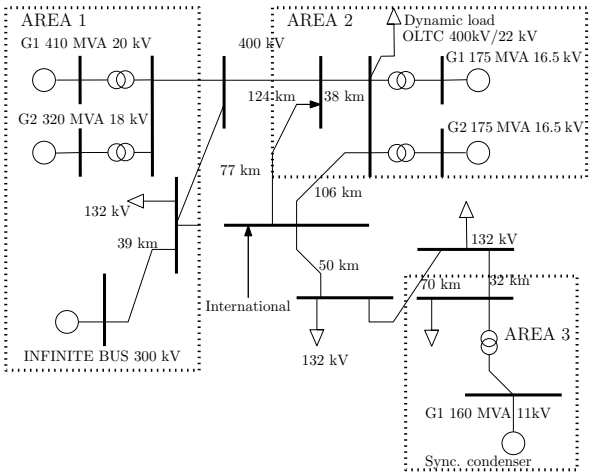


Figure 11. Power system simulator for long term voltage stability simulations based on the Nordic Model from Statnett. The system consists of three different Areas including four synchronous generators and one synchronous condenser. Power is also fed through the international transmission link.

6.3 LTVS Simulations

For LTVS simulations in this paper the international transmission link is disconnected demanding necessary reactive power to be delivered locally. The tap-changer is then trying to restore the voltage at load forcing the system to collapse. As expected from the steady state calculations, the power system have a blackout when part of the system is reaching the critical voltage limit around 0.7 per-unit as seen in Figure 12.

Results from simulations with MPC as primary control are shown in Figure 13 for different tuning parameters. Optimization is done at every $\Delta t = 1s$. A smaller Δt and a longer prediction horizon N_p gives better control but has a

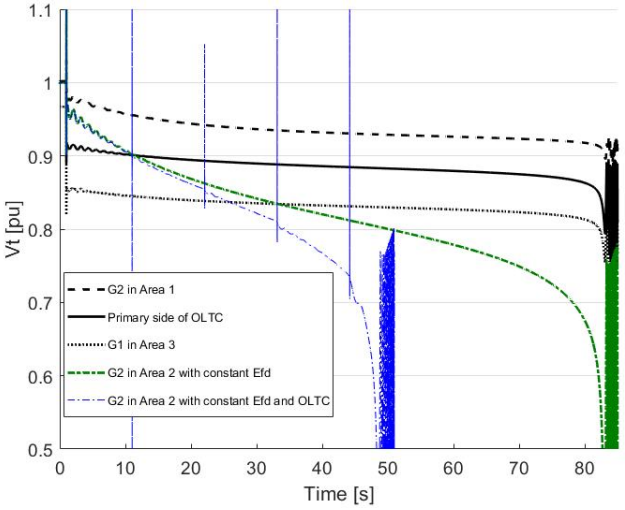


Figure 12. Voltage as a function of time for some selected buses in the power system with and without OLTC control on dynamic load. Area 1 Generator 1 and Generator 2 has constant field voltage. Area 2 Generator 1 and Area 3 Generator 1 has implemented classical control.

higher computational cost. The weighting factors are the same as earlier.

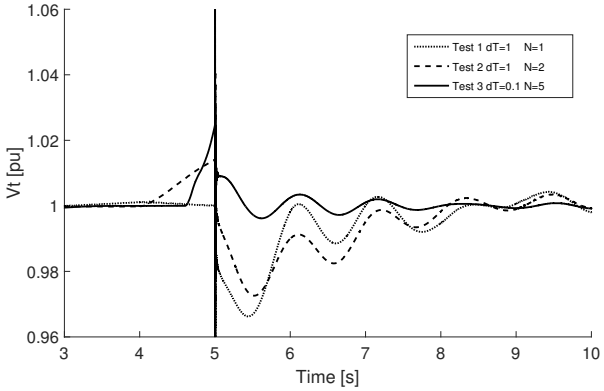


Figure 13. Voltage as a function of time where MPC control action are implemented at Area 2 G2 as primary control.

According to the TSO, the set-point voltage of the controller has to be between 0.9 and 1.05 per-unit value and can be set both locally and from a control center. For simulating such a scenario, MPC is used to change the set-point value $u = V_{ref}$. In this case, the set-point was changed for both generators in Area 2 for bringing the voltage at the On-Load-Tap Controller (OLTC) bus closer to 1 pu after the disturbance. In this simulation, saturation is also included in the dynamic model. Future disturbance is also known. Results are shown in Figure 14 and Figure 15 with different Δt, horizon N_p and tuning parameters.

7 Discussion

This investigation on transient and long term voltage stability considers the ideal (though unrealistic) situation where the internal MPC model exactly matches the real system (perfect MPC model). It is unrealistic to expect that the MPC controller could maintain a complete, accurate system representation. The degree to which MPC can tolerate model inaccuracy is core to practical power system implementation (Gong, 2008). A more realistic implementation could be an SMIB model representing the whole grid. Thus, state estimation should be done on the synchronous generator. In addition, the transfer reactance x_s in (9) needs to be estimated and the infinite bus voltage V_{bus} in (9) needs to adapt the actual voltage level in the system for a good predictive control action. The size of the inductor could be identified from the electromechanical first swing oscillations after a fault.

In these simulations, the full nonlinear model was used with MPC e.g. NMPC. NMPC requires extensive computing power to solve nonlinear constrained optimization problems in real time. In the case of transient stability, NMPC is too slow to avoid loss of synchronism. One other problem that occurs in simulating electrical circuits is that their equations often exhibit stiffness. In SimPowerSystems snubber resistors and capacitors are used across the switches to improve numerical stability. Ode 23tb was the fastest integrator in these simulations. However, with long term voltage stability when the MPC is acting only as set-point changing controller with a sampling time e.g every 5-10 second, NMPC could be feasible. In that case, a linear prediction model combined with C-code implementation should be used.

The MPC should also be tested for low frequency electromechanical oscillations in large interconnected power systems (e.g. both inter-area and local oscillations) (Kundur, 1994). This would determine more precise MPC tuning parameters applicable for both small- and large-signal stability.

The Simscape language is an object-oriented language based on MATLAB and is very attractive to use for power system modeling. One alternative is to integrate the open source object oriented Modelica modeling language with MATLAB through the FMI toolbox from Modelon. In that case the MATLAB optimization toolbox can be used to run MPC on a Modelica model through C code generation.

8 Conclusion

This paper investigates the use of Model Predictive Controll (MPC) for voltage (excitation) control of synchronous generators to enhance the stability of the power system. Simulation results show that a well tuned predictive controller combined with an internal model that exactly matches the real system, gives improved control action in all the simulations compared to classical control. Due to computational limitations, real time simulation for transient stability would not be possible with the strategy

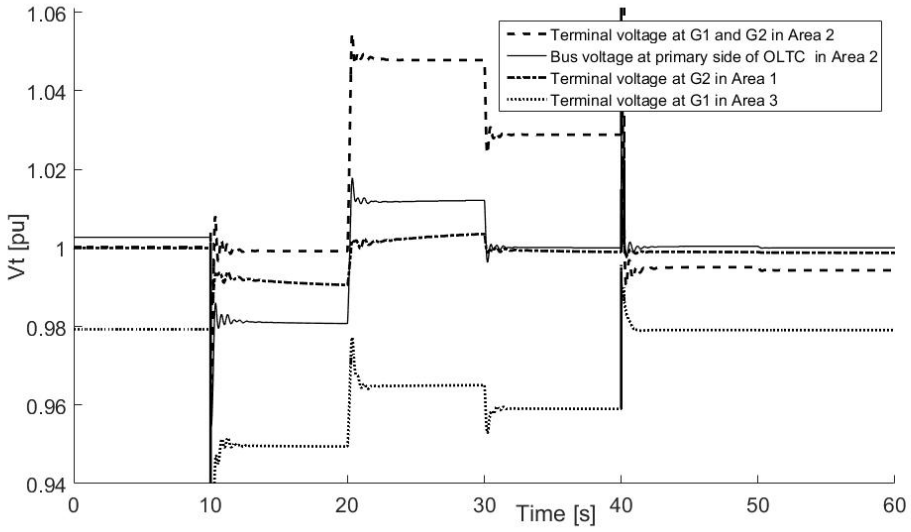


Figure 14. Voltage as a function of time where MPC is changing the set-point on classical control at G1 and G2 in Area 2 to bring voltage at primary side of OLTC bus closer to 1 pu. Optimization is done at every $\Delta t = 10s$ and the horizon is $N_p = 1$. The international transmission line in Figure 11 are disconnected at 10 seconds and reconnected at 40 seconds in this simulations. The MPC tuning parameters were $\mathbf{Q} = 300$, $\mathbf{P} = 5$, $\mathbf{R} = 0.01$.

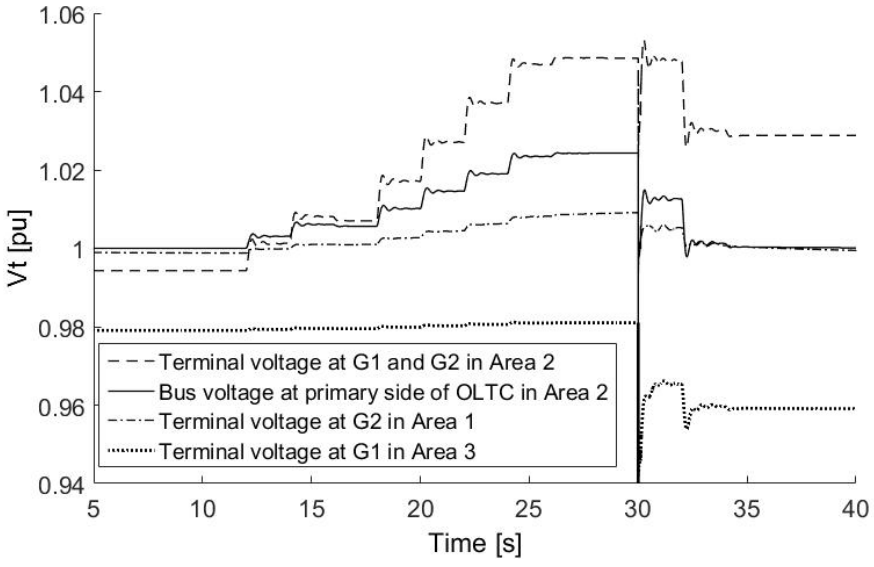


Figure 15. Voltage as a function of time where MPC is changing the set-point on classical control at G1 and G2 in Area2 to bring voltage at primary side of OLTC bus closer to 1 pu. Optimization is done at every $\Delta t = 2s$ and the horizon is $N_p = 10$. The international transmission line in Figure 11 are disconnected at 30 seconds in this simulations. The MPC tuning parameters were $\mathbf{Q} = 30$, $\mathbf{P} = 1$, $\mathbf{R} = 1$.

presented in this paper. However, a reduced power system model combined with C-code generation could be a feasible solution for faster NMPC action.

Acknowledgment

The financial support from Statkraft ASA of the PhD study of the first author is greatly acknowledged. The practical support from Jan Petter Haugli, Statkraft ASA is likewise acknowledged.

References

T. V. Cutsem and C. Vournas. *Voltage Stability of Electric Power Systems*, volume 441. Springer US, 1998.

S. B. Farnham and R. W. Swarthout. Field excitation in relation to machine and system operation [includes discussion]. *Transactions of the American Institute of Electrical Engineers. Part III: Power Apparatus and Systems*, 72(2), Jan 1953. ISSN 0097-2460. doi:10.1109/AIEEPAS.1953.4498759.

B. Gong. *Voltage stability enhancement via model predictive control.* PhD thesis, The University of Wisconsin, Madison, 2008. ProQuest Dissertations and Theses, 173.

J. Grainger and W.D Stevenson. *Power system analysis*. McGraw-Hill, 1994.

G. J. Hegglid. An adaptive multivariable control system for hydroelectric generating units. *Modeling, Identification and Control*, 4(2):63, 1983.

IEEE. Standard definitions for excitation systems for synchronous machines. *IEEE Std 421.1-2007*, pages 1–33, July 2007.

P. Kundur. *Power System Stability and Control*. McGraw-Hill Professional, 1994. ISBN 007035958X.

M. Larsson. *Coordinated Voltage Control in Electric Power Systems*. phdthesis, Lund University, 2000.

J. M. Maciejowski. *Predictive control with constraints*. Pearson education, 2002.

R. Sharma. *Predictive control with implementation, lecture notes.* Telemark University College. Porsgrunn, Norway., 2014.

Statnett. *FIKS Funksjonskrav i Kraftsystemet/Functional requirements in the power system.* Technical report, Statnett, 2012.

T. Øyvang, D. Winkler, B. Lie, and G.J. Hegglid. Power system stability using modelica. In *Proceedings of the 55th Conference on Simulation and Modelling (SIMS 55), Modelling, Simulation and Optimization, 21-22 October 2014, Aalborg, Denmark*, number 108, pages 120–127. Linköping University Electronic Press, Linköpings universitet, 2014.

Modelling and Dynamic Simulation of Cyclically Operated Pulverized Coal-Fired Power Plant

Juha Kuronen Miika Hotti Sami Tuuri
Fortum Power and Heat Oy, Espoo, Finland, {juha.kuronen, miika.hotti, sami.tuuri}@fortum.com

Abstract

Pulverized coal-fired power plants are increasingly operated cyclically for the compensation of fluctuating load in the electric grid caused by intermittent production of wind and solar power plants. Dynamic simulation is a powerful tool for investigating the transient behavior of a power plant that is operated cyclically. New solutions, e.g. process changes and control strategies can be tested with a computer-aided dynamic simulation software. This paper introduces a model and dynamic model validation of a commercial pulverized coal-fired power plant that is used in grid load compensation. The model is developed using dynamic simulation software Apros, and it includes all the main processes and control loops of the plant, but also some significant simplifications have been made compared to the real plant process. The model is validated against measurement data from the plant. Simulated dynamic validation cases include typical load changes that are used in grid load compensation. The model can be used for further investigations regarding flexibility and controllability of the plant.

Keywords: modelling, dynamic simulation, cyclic operation, pulverized coal-fired power plant, validation

1 Introduction

The production of renewable energy, especially wind and solar power, has expanded in the past few years in Europe due to growing concern about the climate change, emission restrictions on the energy markets and renewable energy subsidies. The load level of solar and wind power plants depends strongly on weather conditions which vary depending on the season and daytime. The intermittent renewable energy production leads to fluctuating load in the electric grid, and to balance the grid load some power plants are forced to compensate the load fluctuations by continuously controlling the power output of the plant. This operation mode is called cycling. In Central Europe conventional steam power plants, typically pulverized coal-fired units, are participating in the compensation. Traditionally large coal-fired units have been designed for base load operation, but due to increased renewable energy production base load plants are more and more operated in varying load levels. This sets new challenges for the flexibility and controllability of these plants.

These challenges can be divided into three categories. Firstly faster load transients between operational points as well as faster and more flexible start-up and shutdown processes are needed. Secondly the plant needs to be operated on broader range and the technical minimal load limit has to be re-evaluated. Thirdly the thermo-economical optimization of the plant within the whole operational range need to be done, since the operation in full load is reduced. (Starkloff et al., 2015) In addition the rates of load transients, start-ups and shutdowns are limited by thermo-mechanical stress in boiler and turbine components. Uncontrollable stress reduces significantly the residual life of stress-prone plant components.

In order to respond to the load transient, flexibility, and thermo-economical optimization challenges, development and re-engineering work is typically needed in plants that are not designed for cyclic operation. This work may result e.g. in boiler heat surface constructional changes, burner upgrades or control strategy changes. New solutions can be easily tested with a computer-aided dynamic simulation tool. Hence modelling and dynamic simulation of pulverized coal-fired plants are increasingly important, and this topic has been covered in several recently published papers, e.g. (Starkloff et al., 2015; Richter et al., 2015; Krüger et al., 2015). This paper presents a dynamic simulation model of a commercial 750 MWe pulverized coal-fired power plant that is operated both in base load and cyclic operation mode. The model is validated against operational plant data during typical load changes of the plant.

It is always necessary to know the purpose of the model and the accuracy needed from the model. In this work the main purpose is to simulate load-varying operation of the plant. For this purpose the main process sections and control loops of the plant are included in the model. Water and steam temperatures, pressures and mass flows in the boiler should equate with the real process, so that the load change transients can be simulated accurately. Therefore the boiler section must be modelled carefully. Since a development of an exact plant simulator was not the aim of this work, the model

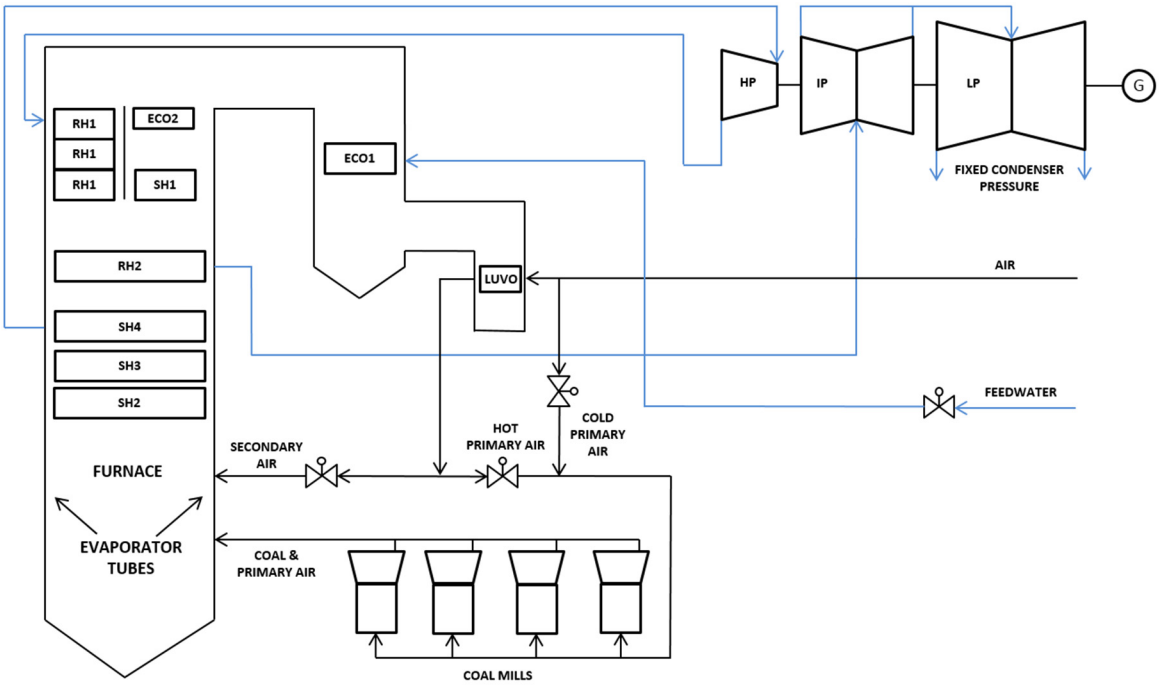


Figure 1. Schematic diagram of the plant model.

was adapted to the specific needs of the simulation experiments. In other words the process and control system models include considered simplifications compared to the real plant. For example feedwater section, including feedwater tank, pumps and preheaters, is not included in the model. The power plant process and control system were modelled in dynamic simulation software Apros.

The paper is structured as follows. After the introduction the simulation software Apros is introduced briefly in Section 2. The plant model is presented in Section 3. The validation results are discussed in Section 4 and the conclusions are given in Section 5.

2 Simulation Environment

Dynamic simulation software Apros was utilized for the modelling and simulation of the reference plant. Apros is a multifunctional software for full-scale modelling and dynamic simulation of power plants and industrial processes. Apros is the result of a quarter century's development work by VTT (Technical Research Centre of Finland) in co-operation with Fortum. It is used by multiple power plant operators, engineering companies, research institutes, safety authorities and universities all over the world.

Apros combines accurate physical process modelling with automation modelling. With Apros it is easy to design, test and see how the process and the control system work together, and the whole integrated system can be studied and optimized simultaneously in detail. The user has access to a set of predefined process component models that are conceptually analogous with concrete devices. The component libraries cover a comprehensive set of process and automation components such as pipes, valves, pumps, heat exchangers, tanks, measurements, controllers etc. One convenient feature in Apros is user component (UC) which can be utilized for creating own re-usable component structures. User component consists of basic Apros library components and possibly other user components. It can be easily re-used and shared with other Apros users.

In Apros, modelling is based on thermal hydraulics which is described using time-dependent conservation equations for mass, momentum and energy as well as correlations for friction and heat transfer (Technical Research Centre of Finland).

3 Plant Model

The reference plant is a commercial pulverized coal-fired unit. The plant is rather new and it represents state of the art technology. It is used both in base load and cyclic operation mode. The once-through boiler of the plant is ultra-supercritical with live steam temperature of 600 °C and live steam pressure of 290 bar. The gross power capacity of the plant is 800 MW and the net electric power is 750 MWe. The plant is also able to produce 90 MWth district heat for the surrounding region. The overview and scope of the plant model is illustrated in Figure 1. The following chapters introduce the plant model. The process is divided into three

process sections and the control system comprises of seven sections.

The simulation model construction and validation can be divided into four steps. Firstly the model scope and boundary conditions are considered and decided based on the use of the model. Then the actual model is built up in the modelling environment. After model build-up the model is tuned to various steady state operating points and the model parameters are tuned by comparing the model responses and plant measurement data. Lastly the dynamic validation is performed by simulating transients that are compared to measurement data. Once these steps are completed successfully the model is ready to be used in different types of simulation experiments.

3.1 Air and Fuel Feed

Air and fuel feed section includes air supply lines, air preheaters, air flow control valves, raw coal supply line, coal mills and coal-air line to the furnace as it is presented in Figure 2.

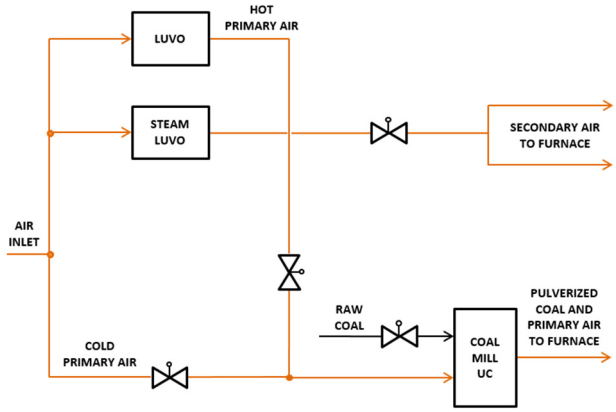


Figure 2. Air and fuel feed section model diagram.

In the model air is passed to the furnace in two phases, whereas in the real plant the number of air feed lines is larger. Primary air, which is fed through the coal mills, makes up about 30 % of all combustion air. Primary air is a mixture of cold and preheated hot air. Hot air is preheated with flue gases in the air preheater, and cold and hot air are mixed at the inlet of the coal mill to control the coal mill outlet temperature. Secondary air makes up the major part of the combustion air, around 70 %. Secondary air is preheated with extraction steam and mixed with fuel-rich primary air in the furnace to give a proper air-fuel ratio.

The coal mills of the plant are modelled in detail using the user component feature of Apros. Coal mill model is an important part of cyclically operated pulverized coal-fired power plant simulator, and a valid mill model improves the accuracy of the plant's transient simulation. Load changes, start-ups and shutdowns can be simulated realistically if the dynamics of the mill model corresponds to dynamics of a real mill. The roll-type coal mill operation is as follows. Raw coal is dropped from the feeder belt into the mill where it lands on the grinding table and is pulverized by rollers. Primary air is used as carrier gas for grinded coal particles. Inside the mill primary air, which is blown from the bottom of the mill, dries the moist coal, picks up the pulverized coal particles from the grinding table and transports particles to the classifier section. Only the finest coal particles escape the mill through the classifier, whereas heavier particles fall back to the grinding table. By modifying the rotation speed of the classifier the amount of escaping coal powder can be controlled.

The proposed coal mill model was first introduced by (Niemczyk et al., 2012). The coal mill UC is tuned to simulate the behavior of the four mills of the real plant. The model is a so-called graybox-model based on physical knowledge and parameter identification methods. Primary air flow and raw coal flow including moisture are the inputs of the mill component. The output is the mixed air-coal flow which is passed to the furnace. In the coal mill model two particle sizes, raw and pulverized coal are considered, and the model comprises of equations for mass balances of raw and pulverized coal in different sections of the mill, overall energy balance and pressure losses inside the mill. In the model the coal storage inside the mill is divided into three different sections: raw and pulverized coal on the grinding table and pulverized coal in the air. The dynamic behavior of the coal storage inside the mill is crucial during load changes. A schematic representation of the mill model is presented in Figure 3. A more accurate description of the mill model is given in (Niemczyk et al., 2012; Kuronen, 2015).

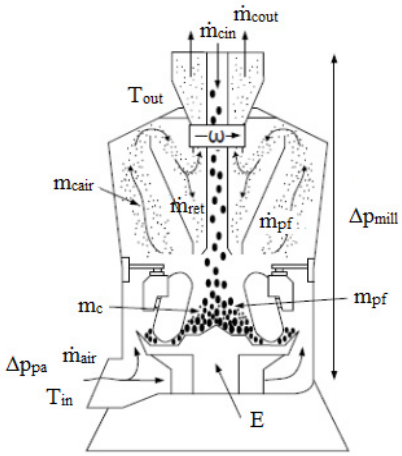


Figure 3. Schematic picture of a roll mill.

3.2 Boiler

The structure of the boiler model is illustrated in Figure 4. In the once-through boiler model the heat transfer between the combustion flames/flue gases and water-steam circuit is modelled by including all the real heat exchange surfaces into the model. These are

economizers, evaporator, superheaters and reheaters. In the real plant the lower part of the evaporator comprises of spiral tube structure and the upper part is constructed with straight tubes. Hence in the model the evaporator is also divided into two separate parts with different dimensions and heat transfer properties. Heat transfer areas and pipe dimensions are fixed according to plant documentation. The heat transfer coefficients, radiation emissitivies and view factors of the heat exchange surfaces are adjusted empirically. Also the water separator, which separates water and steam before the superheating section, and water sprayers, which are utilized in controlling the temperature of the superheated steam, are modelled.

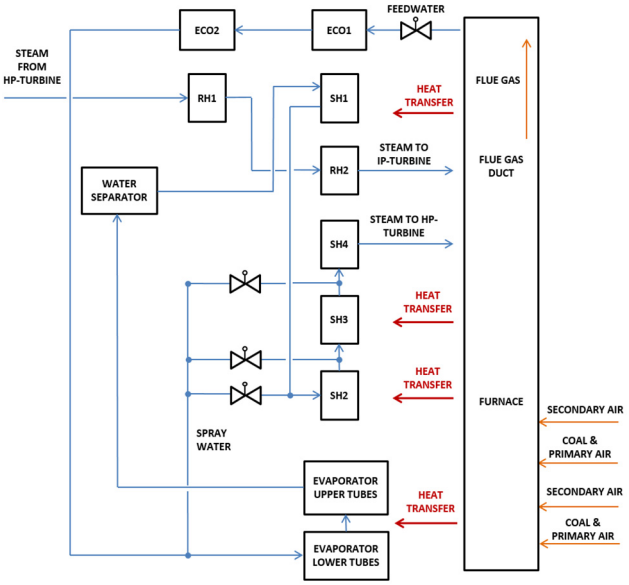


Figure 4. Boiler model diagram.

Flue gas section consists of the combustion chamber, burners and the flue gas duct. The dimensions of the combustion chamber and flue gas duct are equal to ones in the real plant.

3.3 Turbine Plant

Turbine plant model comprises of high-pressure (HP), intermediate-pressure (IP) and low-pressure (LP) turbines, steam extractions, turbine control valve and turbine shaft. Superheated steam flows from the final superheater to the HP-turbine through the turbine control valve, which is kept fully open, since the boiler is operated in sliding pressure mode. After the HP-turbine steam is reheated in two phases in the boiler and brought to IP- and LP-turbines. The mechanical power produced by the turbines is transformed to electric power via shaft and generator components.

The turbine plant model scope does not include a condenser, where the expanded steam is passed after the LP-turbine, but the pressure after the last LP-turbine section is fixed according to the real conditions in the plant condenser. The scheme of the turbine plant model is shown in Figure 5. Turbine sections are composed of turbine components, which are connected to a turbine shaft. After each turbine section, part of the steam is extracted to be taken to the water preheaters and district heat exchangers. Since the feedwater and district heat sections are not included in the model, extraction flows are constantly controlled and the flows alternate according to the operating point. Hence the right mass balance is maintained. Extraction mass flows in different operating points have been collected from the plant's process and information system.

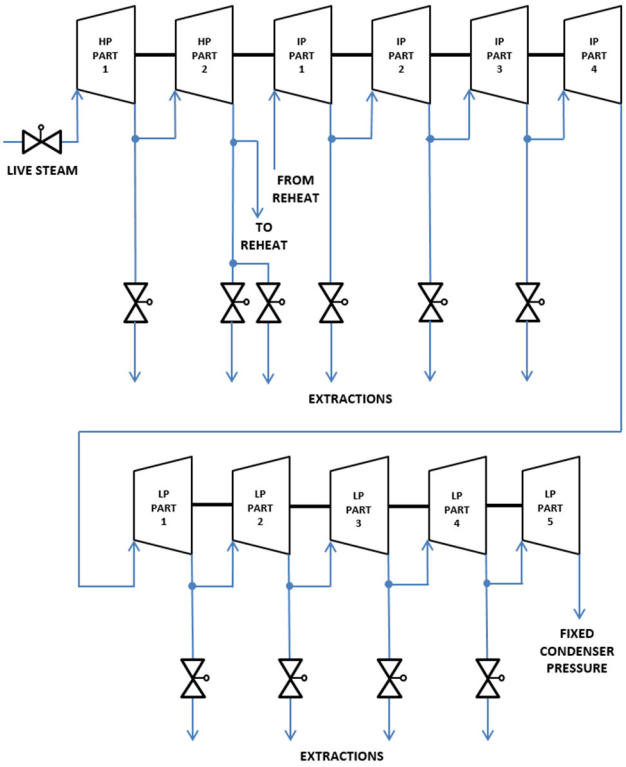


Figure 5. Turbine plant model diagram.

3.4 Control loops

All the main control loops of the plant are modelled and tuned in Apros. The advantage of Apros is that the process and control system can be modelled in the same simulation environment. The modelled control loops are:

- Block control
- Coal flow control
- Primary and secondary air control
- Live steam temperature control
- Reheated steam temperature control
- Feedwater control
- Extraction flow controls.

The co-ordination of turbine and boiler operation is realized by the block control (or unit control), which generates set point for the boiler and turbine to keep the desired load set point while maintaining the desired operating pressures and temperatures (Lamp et al., 2009). The user gives the gross power output set point

via block control and all the other main controlled variables get their set points as a function of the power set point. The user given set point signal is corrected with a power correction controller according to the deviation between the power set point and measured generator power. The block control diagram is presented in Figure 6.

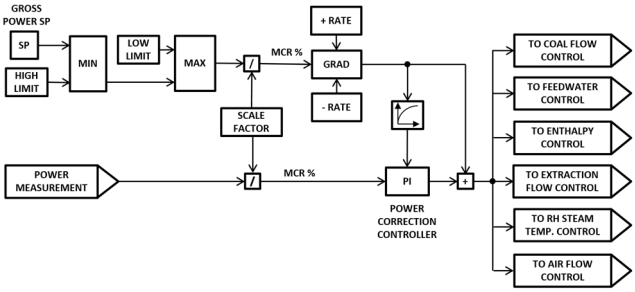


Figure 6. Block control diagram.

The coal flow set point is determined as a function of plant load and it is then adjusted with a heat value correction which takes into consideration the changing content of the raw coal. The coal composition depends on the type of the coal and plants utilize coals from different sources. The changing composition affects strongly on the coal flow control which is based on a function between plant load and coal mass flow. Heat value correction is based on the enthalpy difference between water at the inlet of the boiler and steam at the outlet of the boiler. Designed enthalpy difference is compared with measured one and the coal heat value in the control loop is corrected according to the enthalpy deviation.

The amount of primary air is controlled according to plant load level and coal flow. The mass flow of primary air is approximately two times bigger than coal mass flow. The temperature of the air-coal mixture at the outlet of the mill is controlled with the ratio of preheated and cold primary air. Secondary air flow is adjusted according to the total air demand, which is a function of the power level and coal flow set point. Secondary air set point is modified with O_2-control (oxygen-control), which adjusts the volume of oxygen in the flue gas to the desired level. Secondary air, primary air and coal mill temperature control loops are pieced together in air control diagram.

The live steam temperature is controlled by spraying saturated water among the live steam in three phases. Live steam temperature set point after the last superheater is kept constant in the model. For the first two attemperators the temperature set point is defined by the temperature difference across the attemperator. Reheated steam temperature is controlled correspondingly.

The feedwater control loop is presented in Figure 7. The set point for the feedwater flow is determined according to the load level and it is modified with enthalpy correction. Enthalpy correction factor is calculated according to deviation between the design and measured steam enthalpy after the evaporator. Since steam enthalpy is a function of temperature and pressure, enthalpy correction controls steam temperature and pressure after the evaporator section. Feedwater flow is adjusted with a control valve.

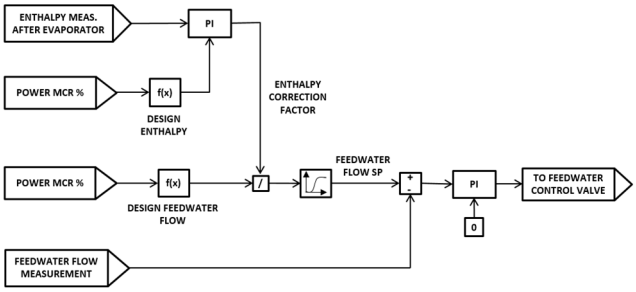


Figure 7. Feedwater control diagram.

Although feedwater preheaters, -tank, -pumps and condenser are not included in the model the extraction flows from the turbines are modelled to achieve the correct mass balance in the water-steam cycle. Extraction flow controls consist of multiple individual valve controls, and set points for the flows are set as a function of plant load level.

4 Model Validation and Simulation Results

The plant model was first validated in multiple steady state conditions between the gross power range of 50-100 %. When the steady state validation results showed good agreement the model was validated dynamically by simulating load changes that are typically performed during grid load compensation. The dynamic validation was done between the gross power range of 87-100 %. The dynamic validation results are presented in this section.

The gross power output is presented in Figure 8. The model response (red line) follows fairly well the measurement curve (blue line), but it does not overshoot or undershoot the set point curve (green line) as strongly as the measurement curve. The deviation between the curves is a consequence of multiple factors, since the inaccuracies of the whole model culminate on the power output response. Real controller tuning parameters and load-dependent set point functions of main variables were not known during the modelling, and it reflects as a deviation between the model and plant responses during load change transients. The high-frequency oscillation in the power curve as well as in feedwater curve is caused by measuring errors and backlash of valves. This type of oscillation is not needed to be included in the model.

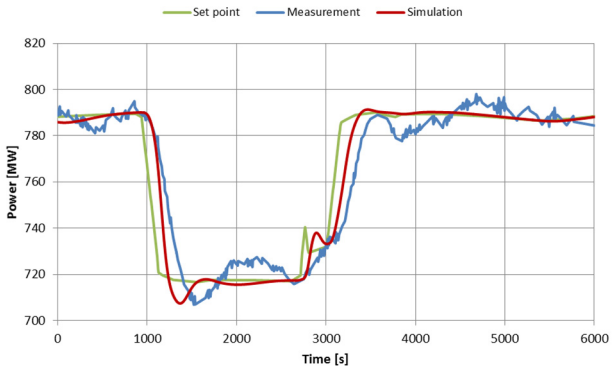


Figure 8. Gross power output.

The raw coal mass flow from the feeder to the coal mill is showed in Figure 9. Dynamically the raw coal flows are quite similar. There are notable peaks in the measured coal flow curve when the load changes are made. The model is able to reproduce the peak when the load is decreased, but there are slight difference between the curves, when the load is increased back to the original level. This originates from the more aggressive controller tuning at the plant. Also the coal flow controller is probably a bit more advanced in the plant control system, whereas in the model an ordinary PI-controller is used. The raw coal flow level is also a bit higher in the model, around 1 kg/s, depending on the operating point. The reason for this can be found from the load-dependent set point function which is a compromise among other variable set point functions. Differences in the coal flow reflect directly on the power output.

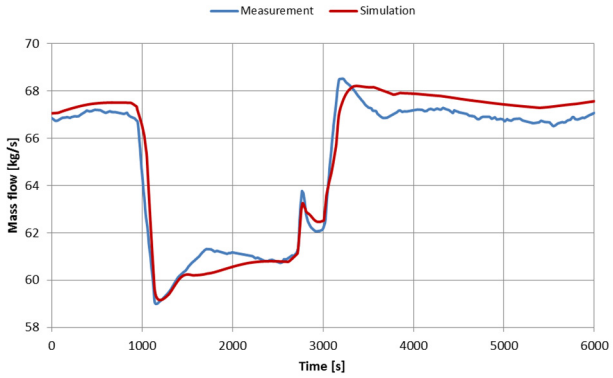


Figure 9. Raw coal mass flow into the coal mill.

The feedwater flow is presented in Figure 10. The simulated feedwater mass flow level is around two percent lower than the measured level. Once again the main reason is the load-dependent set point function which is most probably not the same one that is set in the plant control system. Also the absence of HP- and LP-preheaters, condenser, feedwater tank and the simplified extraction flow modelling have an impact on the feedwater flow in the model. However the curves are dynamically uniform if the high-frequency oscillation is not considered.

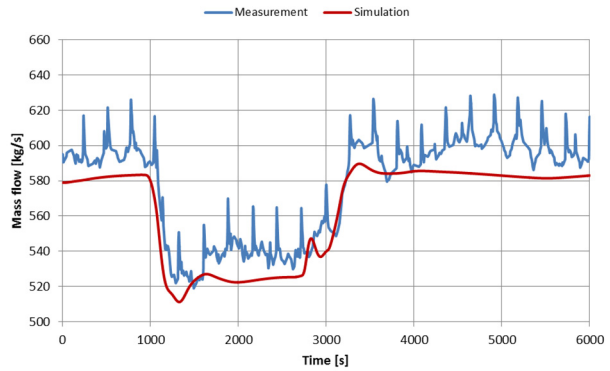


Figure 10. Feedwater mass flow.

The simulated total air mass flow is fairly well in line with the measurement as can be seen from Figure 11. Air controllers at the plant and in the model seem to be tuned quite similarly since the total air mass curves are rather close to each other under the load changes.

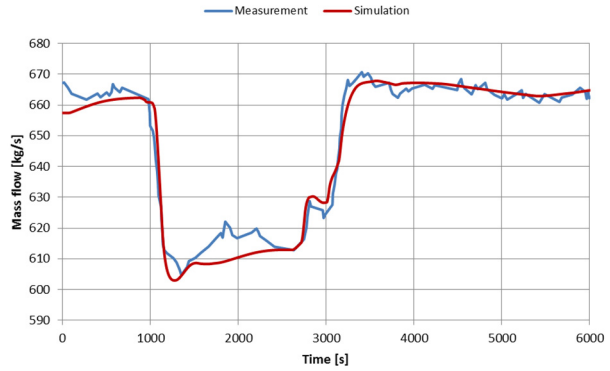


Figure 11. Total air mass flow.

Measured and simulated live steam temperatures after the final superheater are illustrated in Figure 12. The simulated temperature corresponds to the measurement, although the exact same dynamic behavior is difficult to achieve. The correspondence between the live steam temperatures indicates that the dimensions of the heat exchange surfaces and heat transfer coefficients in the boiler are substantially correct. Also the live steam temperature control maintains the temperature close to the set point (597 °C) during the load changes.

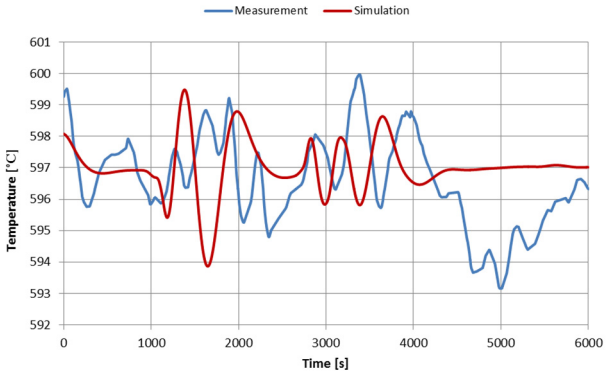


Figure 12. Live steam temperature.

5 Conclusions

Pulverized coal-fired power plants are increasingly participating in electric grid load control due to expanded intermittent renewable power generation. This sets new challenges related to flexibility and controllability of the coal plants since the plants are operated cyclically to maintain a steady load in the electric grid. For example process- and control strategy changes are often needed in base load plants which are not designed to be operated cyclically. New solutions can be designed and tested with time-variant dynamic simulation models.

This paper presented a dynamic simulation model of a full-scale commercial pulverized coal-fired power plant. The model was constructed and simulated in dynamic simulation software Apros. The model included the main processes and control loops of the real plant, but some considered simplifications were done to delimit the scope of the model. The model was validated both on steady-state and dynamic conditions. Based on the dynamic validation results, which were illustrated in this paper, the model responses corresponded fairly well to the plant measurements between the power range 87-100 %. Even if the model is a simplified version of the real plant, it can be already utilized e.g. for design and testing purposes.

In future the plant model can be developed in various ways. Firstly new subprocesses and control loops could be added to the model. Secondly re-tuning the control loops, defining appropriate time constants and load-dependent set point functions would increase the model accuracy. Thirdly a more comprehensive model validation on a broader power range should be performed, so that the model could be reliably simulated on the plant's whole operating range.

References

U. Krüger, M. Rech, S. Tuuri. H. Zindler, Dynamischer Kraftwerkssimulator zur leittechnischen Optimierung der Sekundärantwort des E.ON-Kraftwerks Wilhelmshaven, *Kraftwerkstechnik 2015 - Strategien, Anlagentechnik und Betrieb*, Technische Universität Dresden, pages 641-649, 2015.

J. Kuronen, *Improving Transient Simulation of Pulverized Coal-Fired Power Plants in Dynamic Simulation Software*, Master's Thesis, Tampere University of Technology, 88 p, 2016.

B. Lamp, K. Wendelberger, B. Meerbeck, A New Era in Power Plant Control Performance, Siemens, *Reprint from COAL POWER Magazine*, 6 p, 2009.

P. Niemczyk, P. Andersen, J. D. Bendtsen, T. Sondergaard Pedersen, A. P. Ravn, Derivation and Validation of a Coal Mill Model for Control, *Control Engineering Practice*, 20(5): 519-530, 2012.

M. Richter, F. Möllenbruck, A. Starinski, G. Oeljeklaus, K. Görner, Flexibilization of Coal-Fired Power Plants by Dynamic Simulation, In *Proceedings of the 11th International Modelica Conference*, pages 715-723, 2015.

R. Starkloff, F. Alobaid, K. Karner, B. Epple, M. Schmitz, F. Boehm, Development and Validation of a Dynamic Simulation Model for a Large Coal-Fired Power Plant, *Applied Thermal Engineering*, 91: 496-506, Elsevier, 2015.

Technical Research Centre of Finland, *Apros Thermal Hydraulics*, *Thermal Hydraulic Flow Models*, Available: http://www.apros.fi/filebank/71-Apros_thermal_hydraulics_-_general.pdf.

Hardware-in-the-Loop Emulation of Three-Phase Grid Impedance for characterizing Impedance-Based Instability

Tuomas Messo[1] Jussi Sihvo[1] Tomi Roinila[2] Tommi Reinikka[2] Roni Luhtala[2]

[1]Department of Electrical Engineering, Tampere University of Technology, Finland, {tuomas.messo}@tut.fi
[2]Department of Automation Sciences and Engineering, Tampere University of Technology, Finland

Abstract

The amount of grid-connected power electronic converters is increasing as the world's energy production shifts toward sustainable sources. Poor power quality and harmonic resonances have been reported which have been shown to be caused by grid-connected converters. Accurate modeling tools are required to characterize the conditions for instability and to design stable power-electronics-based power systems.

Unstable behavior can be identified by using models implemented in circuit simulators or using power-hardware-in-the-loop setups. The unstable resonance occurs when inverter control system interacts with the grid impedance. However, a very wide impedance-bank is required in the laboratory to test inverter stability when grid impedance is expected to vary significantly. Moreover, stability tests are often limited to cases where grid impedance is approximated as an inductance. This paper proposes a method for emulating the grid impedance in a hardware-in-the-loop setup which eliminates the need for bulky passive components and allows arbitrary grid impedance to be emulated. As a result, the inverter can be tested with a varying grid impedance to determine the exact conditions for unstable behavior. Moreover, the grid impedance can be changed online to emulate the behavior of a time-varying power grid in real time.

Keywords: hardware-in-the-loop, grid impedance, emulation

1 Introduction

The amount of grid-connected power electronic converters, such as wind and solar inverters is rapidly increasing as the amount of renewable power generation is ramped up (Bose, 2013). Stability of conventional power systems has been mainly determined by power balance, i.e., the produced power has to match the load power to keep grid frequency and amplitude within acceptable limits. However, the stability of modern power systems is challenged by the dynamic behavior of grid-connected power electronic converters, and especially, their small-signal characteristics (Wang et al., 2014; Messo et al., 2013).

Grid-connected converters have been shown to introduce stability issues when the penetration of renewables is significant (Enslin and Heskes, 2004). This has been shown to be mainly caused by poorly damped resonance formed together by inverter output impedance and the grid impedance (Sun, 2011; Wen et al., 2016; Suntio et al., 2017). Wind power inverters were found out to cause instability when connected to a series-compensated line (Belkin, 2010) and photovoltaic inverters have been reported to become unstable in a grid that has high inductance (Yang et al., 2014). Instability and resonance issues can be prevented if the inverter impedance is designed to have larger impedance than the grid or if both impedances resemble passive circuits (Harnefors et al., 2016).

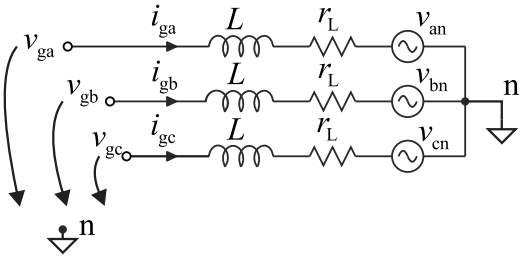


Figure 1. Three-phase grid impedance.

Immunity of a grid-connected inverter to impedance-based instability can be evaluated by connecting the inverter to the power system through an inductance which has sufficiently large value. A weak grid is usually approximated by a large inductance or combination of resistive and inductive elements (Gavrilovic, 1991). However, in reality the grid impedance can be hardly described as a lossy inductor since it may exhibit resonances (Jessen et al., 2015). In practice this means that one should have sufficient physical impedance-bank in the laboratory to test converters in different grid conditions. An interesting alternative is to use a real-time simulator paired with a linear amplifier to emulate the grid impedance which gives more freedom to define the nature of grid impedance, e.g., a series or parallel resonance. Promising results have been presented in the literature, such as in (Kotsampopoulos et al., 2015) where a real-time simulator was used to emulate part of the grid impedance in a hardware-in-the-loop simulation. However, the performance of the impedance emulation method was not validated by measuring the emulated impedance.

This paper proposes a method to emulate the grid

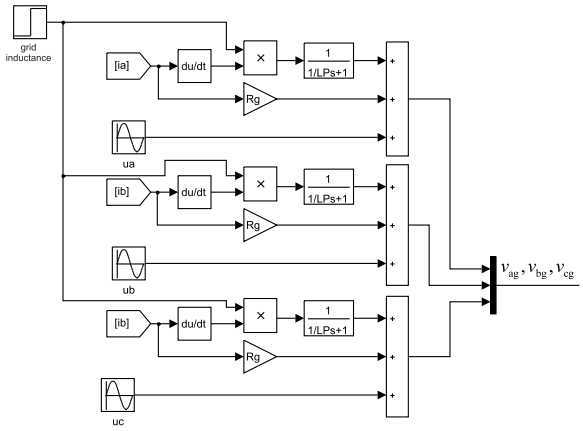


Figure 2. Simulink model to realize the grid impedance in Figure 1.

impedance behavior using power-hardware-in-the-loop (PHIL) setup, which can be used in characterizing impedance-based stability of grid-connected inverters. The method enables modifying the grid impedance online allowing stability studies of grid-connected converters in time-varying grid conditions. The limitations of PHIL-implementation due to sampling delay are also discussed. The delay is shown to introduce a considerable error in the phase of the emulated grid impedance.

The paper is organized as follows: Section II explains how a three-phase grid impedance is modeled in the MATLAB Simulink environment. Such model can be directly build into C-code and ran on a real-time simulator. Section III shows a simulation case where grid inductance is stepped up to destabilize a grid-connected inverter. The practical implementation of the grid impedance emulator and its limitations are discussed in Section IV where a dSPACE real-time simulator and a linear amplifier are used to emulate an inductive grid impedance. Final conclusions are summarized in Section V.

2 Grid impedance model

Circuit diagram of a three-phase grid impedance is as depicted in Figure 1 where v_{an}, v_{bn} and v_{cn} represent the ideal grid voltages. The grid is formed by three identical branches which have resistance and inductance. Grid voltages seen by a grid-connected inverter, i.e., v_{ga}, v_{gb} and v_{gc} at the PCC, can be solved by utilizing basic circuit theory and given according to (1).

$$\begin{aligned} v_{\mathrm{ga}}(t) &= L\frac{\mathrm{d}}{\mathrm{d}t}i_{\mathrm{ga}}(t) + r_{\mathrm{L}}i_{\mathrm{ga}}(t) + v_{\mathrm{an}} \\ v_{\mathrm{gb}}(t) &= L\frac{\mathrm{d}}{\mathrm{d}t}i_{\mathrm{gb}}(t) + r_{\mathrm{L}}i_{\mathrm{gb}}(t) + v_{\mathrm{bn}} \\ v_{\mathrm{gc}}(t) &= L\frac{\mathrm{d}}{\mathrm{d}t}i_{\mathrm{gc}}(t) + r_{\mathrm{L}}i_{\mathrm{gc}}(t) + v_{\mathrm{cn}} \end{aligned} \tag{1}$$

A three-phase grid impedance can be emulated using a linear amplifier where the reference voltages are calculated based on the measured grid currents according to (1). This is an interesting idea since practically the amplifier could be configured to resemble an arbitrary grid impedance which is very attractive for stability studies of grid-connected inverters and for quality control of power converters.

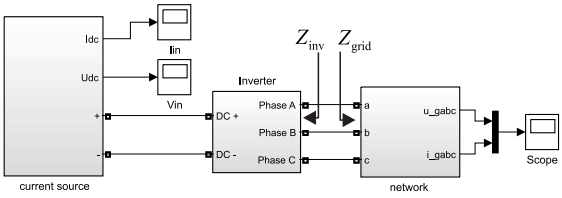


Figure 3. Overview of the grid-connected inverter.

A Simulink model according to to (1) was built to emulate the dynamics of grid impedance, as depicted in Figure 2. Moreover, a low-pass filter with a cut-off frequency of 10 kHz was used to filter out switching ripple from the output of the derivative block. The value of grid inductance can be changed online which emulates the behavior of a real power grid. Moreover, the model could be easily modified to enable online variation of frequency, phase angle and resistance. However, the scope of this paper is limited to changing the value of inductance.

3 Impedance-based instability studies

The grid model was connected to a switching model of a three-phase inverter as illustrated in Figure 3. The model is built using the SimScape component library. The inverter is fed from a DC current source and connected to three-phase voltage sources. The reference values of grid voltages at the PCC were calculated according to (1). The grid currents become unstable when the ratio of inverter output impedance and the simulated grid impedance fail to satisfy the Nyquist stability criterion, i.e., when impedance q-components $Z^{\mathrm{q}}_{\mathrm{inv}}$ and $Z^{\mathrm{q}}_{\mathrm{grid}}$ form together an undamped resonance. Derivation of the inverter impedance model and discussing the stability criterion are out of the scope of this paper but the reader is urged to see (Messo et al., 2015) that shows how the Nyquist stability criterion can be applied to three-phase impedances.

Figure 4 shows the impedance ratio $Z^{\mathrm{q}}_{\mathrm{grid}}/Z^{\mathrm{q}}_{\mathrm{inv}}$ on the complex-plane when grid inductance is selected as 2 mH and increased to 5 mH. The Nyquist stability criterion states that the system is unstable when the impedance ratio encircles the point (-1,0).Thus, the system is stable when grid inductance is 2 mH but becomes unstable when the inductance is increased to 5 mH.

The "Inverter" block in Figure 3 includes the actual power stage, AC filter and the control system. The inverter utilizes normal cascaded control scheme where DC voltage and AC currents are controlled to keep the DC voltage at a certain level and to feed power to the grid at unity power factor. The inverter utilizes a phase-locked-loop

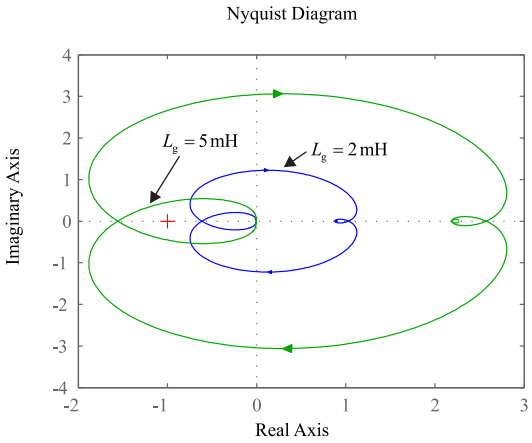


Figure 4. Ratio of inverter impedance and grid impedance on the complex-plane.

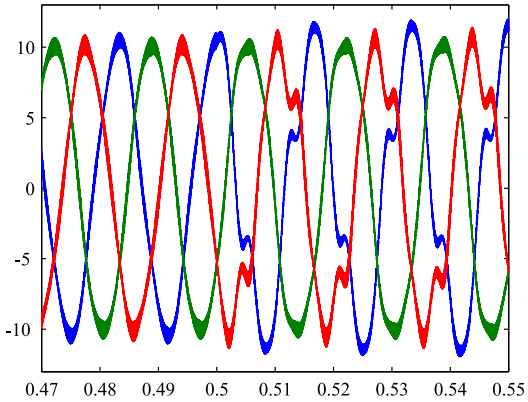


Figure 5. Grid currents become unstable due to increase in grid inductance from 2 to 5 mH at 0.5 s.

(PLL) to synchronize its output currents to grid voltages. The power stage and necessary measurements are built using the components found in the SimScape-toolbox as illustrated in Figure 6.

Phase-locked-loop makes the q-component of inverter output impedance resemble a negative resistance which can destabilize the inverter in a weak grid (Messo et al., 2013). The PLL was tuned to have a crossover frequency of 200 Hz which causes instability when grid inductance increases to 5 mH. Figure 5 shows simulated grid currents when the grid inductance is suddenly increased from 2 to 5 mH at 0.5 s and the inverter becomes unstable due to high-bandwidth PLL. The simulator could be constructed by adding grid impedance in the model as inductive components and connecting an extra 3 mH inductance at 0.5 s by using an ideal switch. However, in this case the simulation time is considerably longer. Therefore, using the proposed impedance model as in Figure 2 also enables smaller simulation time.

4 Practical implementation

The feasibility of grid impedance emulation method in a power-hardware-in-the-loop environment was tested by implementing the impedance model using a dSPACE real-time simulator and a three-phase linear amplifier PAS15000 manufactured by Spitzenberger. Photograph of the laboratory setup is as shown in Figure 7. The setup consists of a three-phase IGBT inverter supplied by a PV simulator feeding the three-phase linear amplifier through an isolation transformer.

The Simulink model for emulating inductive grid impedance is as shown in Figure 8. Ideal grid voltages were generated by three sine-wave sources phase shifted by 120 degrees. Voltage drop over the emulated phase inductance was obtained by multiplying the inductor current derivative by the desired inductance value. The output signal of the product block had to be low-pass filtered to avoid destabilizing the linear amplifier which is due to sampling delay of dSPACE. The impedance model was running on the same dSPACE-platform as the inverter control system. Therefore, the sampling frequency has to be set the same as inverter switching frequency (8-12 kHz).

Input impedance of the linear amplifier, i.e., the emulated grid impedance was measured in the dq-domain using the inverter as a perturbation source and by measuring the frequency response from grid current q-component to grid voltage q-component. The grid impedance q-component is the ratio of these components as given in (2). The measurement setup is as depicted in Figure 9. PRBS-injection was added in the reference value of inverter output current q-component. The grid current and voltage q-components were measured and the grid impedance was computed using the methods discussed in (Roinila et al., 2015). The currents and voltage were measured in the dq-reference frame tied to the inverter control system, i.e., by utilizing the grid voltage angle estimated by the PLL. The PLL crossover frequency was set to 2 Hz to avoid PLL from affecting the measured impedance at low frequencies.

$$Z_g^q = \frac{\hat{v}_{oq}}{\hat{i}_{oq}}. \tag{2}$$

The isolation transformer has some resistive losses and stray inductance which have to be measured first to evaluate which part of the grid impedance is caused by the transformer and which is realized by the impedance emulator. The grid impedance q-component was measured while the value of emulated inductance value was set to zero. The measured and fitted impedances are shown in Figure 10. The isolation transformer has resistance of 400 mΩ and inductance of approximately 600 μH. I.e., the phase is around 0 degrees at low frequencies with a constant magnitude and increases to 90 degrees at higher frequencies with magnitude increasing 20 dB per decade. It can be concluded that the impedance measurement is

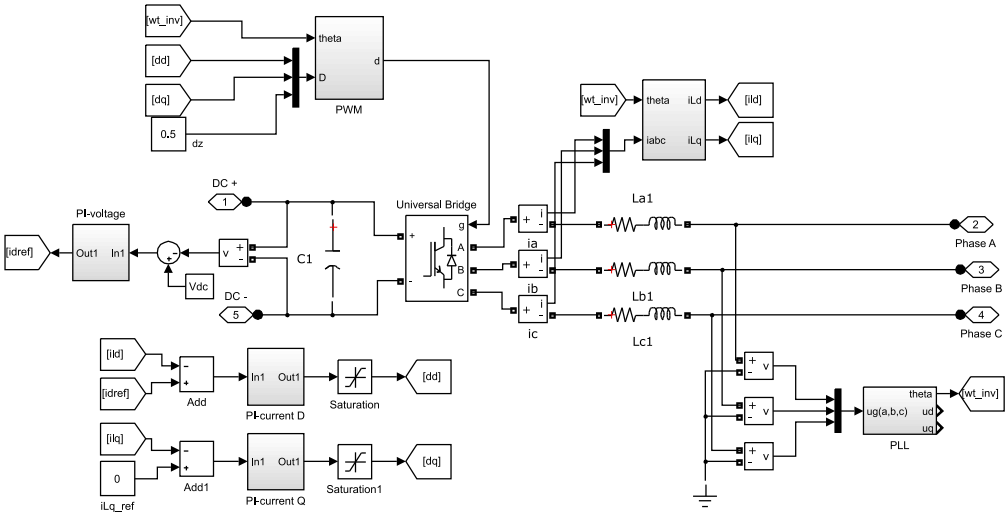


Figure 6. Inverter power stage with control system in Simulink.

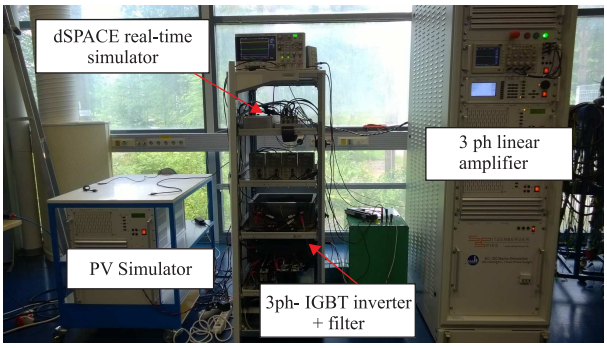


Figure 7. Three-phase inverter connected to a three-phase grid emulator.

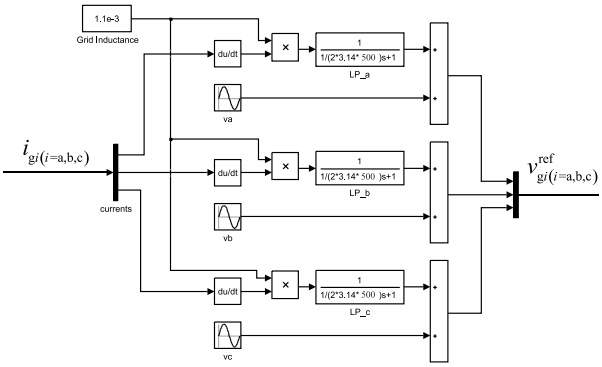


Figure 8. Impedance model implemented using the dSPACE real-time simulator.

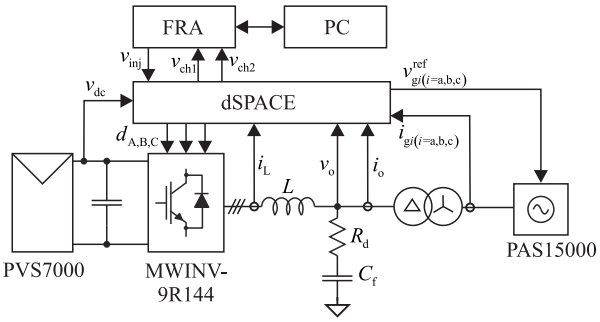


Figure 9. Setup for measuring the emulated grid impedance.

accurate up to few kilohertz after which the phase curve starts to drop due to sampling delay.

Figure 11 shows the measured grid impedance in green when the impedance emulator is activated. The inductance value was set to 1.1 mH and the low-pass filter was tuned to have a crossover frequency of 500 Hz allowing stable operation of the linear amplifier. The red dots illustrate the impedance of a series RL-circuit with resistance equal to 400 mΩ and inductance of 1.7 mH, i.e., the reference grid impedance includes the effect of isolation transformer. The magnitude of the emulated impedance follows the reference curve up to 400 Hz after which the impedance experiences an additional series resonance. Moreover, the phase starts to deviate from the reference value already around 50 Hz. The inaccuracy of emulated impedance is caused by the low-pass filter which was required for stable operation. A low-pass filter with a single pole starts decreasing the phase of the impedance already one decade below cut-off frequency, i.e., at 50 Hz. However, the low-pass filter is required for stability and cannot be set to higher frequency while sampling frequency is tied to inverter switching frequency. Reference of the grid impedance which includes the effect of transformer

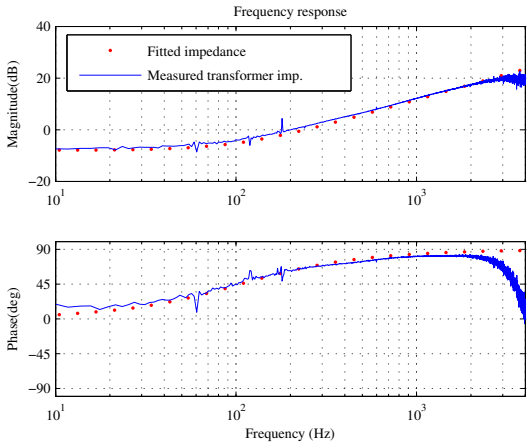


Figure 10. Measured transformer impedance q-component.

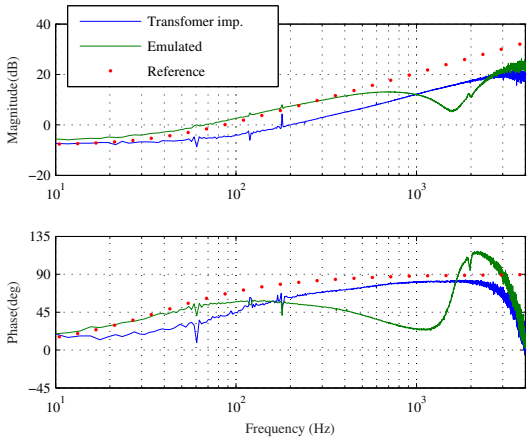


Figure 11. Measured grid impedance when impedance emulator is activated.

impedance is given as a function of frequency in (3).

$$Z_{\mathrm{g}}^{\mathrm{q}}(j\omega) = 0.4 + j\omega 1.7 \cdot 10^{-3} \tag{3}$$

Switching frequency of the inverter was increased to 12 kHz and, therefore, the sampling frequency of the grid impedance emulator was increased as well. The sampling delay was effectively reduced by a factor of 1.5. The higher sampling frequency allows stable operation of the linear amplifier with emulated inductance value of 900 μH when the low-pass filter was tuned to have cut-off frequency of 1 kHz. Using higher value of emulated inductance value or increasing the cut-off frequency would destabilize the linear amplifier. This should be avoided since very large high frequency currents would flow in the circuit. (Enough to trip current sensor with 50 A current limit!)

Measured grid impedance and its reference value are shown in Figure 12. The magnitude curve follows the reference curve up to almost 1 kHz. However, the phase curve begins to deviate from reference curve approximately after 100 Hz. This is expected since the low-pass

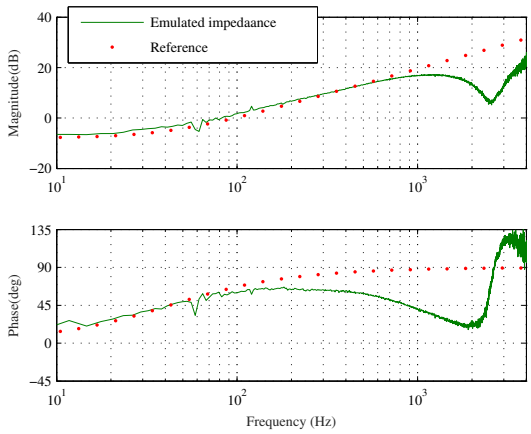


Figure 12. Measured grid impedance when impedance emulator is activated with higher sampling frequency.

filter has cut-off frequency of 1 kHz. It is not reasonable to increase the inverter switching frequency much higher since the IGBT switches require as much as 4 μs of blanking time to avoid shorting the DC capacitor. The reference curve for the emulated impedance is calculated according to (4).

$$Z_{\mathrm{g}}^{\mathrm{q}}(j\omega) = 0.4 + j\omega 1.5 \cdot 10^{-3} \tag{4}$$

Based on the measured impedances it can be concluded that the present impedance emulator is suitable for characterizing impedance-based instabilities occuring at low-frequencies in cases when grid inductance has a maximum value around 1 mH. The limitation arises from the low sampling frequency which causes the phase of emulated impedance to deviate from the reference value at frequencies higher than 100 Hz. Impedance-based interactions occurring at higher frequencies cannot be, therefore, reproduced. As a future research the sampling frequency of impedance emulator should be decoupled from the inverter control system which requires another real-time simulator for implementation.

5 Conclusions

The amount of grid-connected power electronic converters is increasing due to reduced price of renewable energy, such as wind and solar. At the same time stability of power grids is challenged by the inverter control functions, such as grid-synchronization algorithms. Evaluation methods to characterize stability issues introduced by grid-connected converters need to be developed to enable large-scale utilization of renewable energy.

The converter becomes unstable when its control system starts interacting with grid impedance which usually varies over time. In determining impedance-based stability, a large set of physical components is required in the laboratory to realize the variation in grid impedance. Moreover, contactors are required to change the grid

impedance to emulate load changes in the grid which complicates the test setup and increases cost.

This paper studies a grid impedance emulation method using a dSPACE real-time simulator and three-phase linear amplifier which eliminates the need for bulky and expensive passive components. In the proposed method grid impedance is modeled inside the real-time simulator which allows changing the value of grid impedance online. Performance of the impedance emulation method is studied by measuring the input impedance of the impedance-emulator by using a three-phase IGBT inverter to generate the small-signal excitation in its output currents. The impedance emulator is shown to be very sensitive to sampling delay which degrades the phase behavior of the emulated grid impedance. Moreover, the delay easily destabilizes the linear amplifier when large grid inductance is to be emulated. This issue should be treated with caution since very large high-frequency currents can flow in the circuit. Future work will include implementing the grid impedance emulation method by an additional real-time simulator in order to achieve higher sampling frequency and smaller delay.

6 Aknowledgements

This research was supported by the Academy of Finland.

References

P. Belkin. Event of 10-22-09. *CREZ Technical Conference, Electrical Reliability Council of Texas*, 2010. URL `www.ercot.com`.

B. K. Bose. Global energy scenario and impact of power electronics in 21st century. *IEEE Trans. Ind. Electron.*, 60(7): 2638–2651, 2013. doi:10.1109/TIE.2012.2203771.

J. H. R. Enslin and P. J. M. Heskes. Harmonic interaction between a large number of distributed power inverters and the distribution network. *IEEE Trans. Power Electron.*, 19(6): 1586–1593, 2004. doi:10.1109/TPEL.2004.836615.

A. Gavrilovic. Ac/dc system strength as indicated by short circuit ratios. *International Conference on AC and DC Power Transmission*, pages 27–32, 1991.

L. Harnefors, X. Wang, A. Yepes, and F. Blaabjerg. Passivity-based stability assessment of grid-connected vscs - an overview. *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 4(1):116–125, 2016. doi:10.1109/JESTPE.2015.2490549.

L. Jessen, F. W. Fuchs, and C. Kiel. Modeling of inverter output impedance for stability analysis in combination with measured grid impedances. *IEEE 6th Int. Symp. Power Electron. Distrib. Gener. Syst. (PEDG)*, pages 1–7, 2015. doi:10.1109/PEDG.2015.7223037.

P. Kotsampopoulos, F. Lehfuss, G. Lauss, B. Bletterie, and N. Hatziargyriou. The limitations of digital simulation and the advantages of phil testing in studying distributed generation provision of ancillary services. *IEEE Trans. Ind. Electron.*, 62(9):5502–5515, 2015. doi:10.1109/TIE.2015.2414899.

T. Messo, J. Jokipii, A. Mäkinen, and T. Suntio. Modeling the grid synchronization induced negative-resistor-like behavior in the output impedance of a three-phase photovoltaic inverter. *4th IEEE International Symposium on Power Electronics for Distributed Generation Systems (PEDG)*, pages 1–7, 2013. doi:10.1109/PEDG.2013.6785602.

T. Messo, A. Aapro, and T. Suntio. Generalized multivariable small-signal model of three-phase grid-connected inverter in dq-domain. *IEEE 16th Workshop on Control and Modeling for Power Electronics (COMPEL)*, pages 1–8, 2015. doi:10.1109/COMPEL.2015.7236460.

T. Roinila, T. Messo, T. Suntio, and M. Vilkko. Pseudo-random sequences in DQ-domain analysis of feedforward control in grid-connected inverters. *IFAC-PapersOnLine*, 48(28):1301–1306, 2015. doi:10.1016/j.ifacol.2015.12.311.

J. Sun. Impedance-based stability criterion for grid-connected inverters. *IEEE Trans. Power Electron.*, 26(11):3075–3078, 2011. doi:10.1109/TPEL.2011.2136439.

Teuvo Suntio, Tuomas Messo, and Joonas Puukko. *Power Electronic Converters: Dynamics and Control in Conventional and Renewable Energy Applications*. Wiley VCH, 2017. ISBN 978-3-527-34022-4.

X. Wang, F. Blaabjerg, and W. Wu. Modeling and analysis of harmonic stability in an ac power-electronics-based power system. *IEEE Trans. Power Electron.*, 29(12):6421–6432, 2014. doi:10.1109/TPEL.2014.2306432.

B. Wen, D. Boroyevich, R. Burgos, P. Mattavelli, and Z. Shen. Analysis of d-q small-signal impedance of grid-tied inverters. *IEEE Trans. Power Electron.*, 31(1):675–687, 2016. doi:10.1109/TPEL.2015.2398192.

D. Yang, X. Ruan, and H. Wu. Impedance shaping of the grid-connected inverter with lcl filter to improve its adaptability to the weak grid condition. *IEEE Trans. Power Electron.*, 29 (11):5795–5805, 2014. doi:10.1109/TPEL.2014.2300235.

Parametric CFD Analysis to study the Influence of Fin Geometry on the Performance of a Fin and Tube Heat Exchanger

Shobhana Singh, Kim Sørensen, Thomas J. Condra

Department of Energy Technology, Pontoppidanstræde 9220, Aalborg East, Denmark,
ssi@et.aau.dk; kso@et.aau.dk; tc@et.aau.dk

Abstract

Heat transfer and pressure loss characteristics of a fin and tube heat exchanger are numerically investigated based on parametric fin geometry. The cross-flow type heat exchanger with circular tubes and rectangular fin profile is selected as a reference design. The fin geometry is varied using a design aspect ratio as a variable parameter in a range of 0.1-1.0 to predict the impact on overall performance of the heat exchanger. In this paper, geometric profiles with a constant thickness of fin base are studied. Three-dimensional, steady-state CFD model is developed using commercially available Multiphysics software COMSOL v5.2. The numerical results are obtained for Reynolds number in a range from 5000 to 13000 and verified with the experimentally developed correlations. Dimensionless performance parameters such as Nusselt number, Euler number, efficiency index, and area-goodness factor are determined. The best performed geometric fin profile based on the higher heat transfer and lower pressure loss is predicted. The study provides insights into the impact of fin geometry on the heat transfer performance that help escalate the understanding of heat exchanger designing and manufacturing at a minimum cost.

Keywords: fin and tube heat exchanger, numerical modelling, fin profile, conjugate heat transfer, turbulent flow, pressure loss

1 Introduction

Fins are the extended surfaces used in heat exchangers to enhance the heat transfer rate between heat transfer surfaces and the flowing fluid (Cengel et al., 2012). The increment in the heat transfer performance through fin surfaces is widely employed in many industrial applications. Application of waste heat recovery systems has received tremendous attention during the last decade due to the resulting saving of primary fuel, increased energy efficiency and lower greenhouse gas emissions. Heat exchangers are one of the important components of these waste heat recovery systems. During past few years, H-type finned and tube heat exchangers have been studied both experimentally (Yu et al., 2010; Chen and Lai, 2012; Chen et al., 2014) and numerically (Tong, 2007; Zhang et al., 2010; Jin et al., 2013). The studies mainly focused on examining the heat transfer and flow resistance characteristics for a reference design of the H-type finned tube bundles. In addition, combined heat and mass transfer analysis on H-type design with three types of finned tube namely-dimple finned tube, longitudinal vortex generators (LVGs) finned tube, and finned tube with compound dimples and LVGs together was conducted (Wang and Tang, 2014; Zhao et al.,2014).

The implementation of fins on the primary heat surface enhances the complexity, volume, and weight which make the design and construction of fin surfaces of vital importance in heat exchanger applications. Very limited research on different fin types or geometry profiles is available due to restricted experimental conditions and numerical challenges. This limitation overshadows the current knowledge of design factors that influence the heat transfer and pressure loss characteristics. Hence, it becomes imperative to study the different fin geometric profiles to determine the optimal fin design for a given H-type fin and tube heat exchanger application.

In this paper, we used Computational Fluid Dynamics (CFD) to obtain the solution of governing equations of physical phenomena in a cross-flow type fin and tube heat exchanger. The parametric study of fin geometry is conducted using air as a working fluid considering the 'rectangular' fin as reference geometric profile. Heat transfer and pressure loss characteristics in a fin and tube heat exchanger with different geometric fin profiles are predicted and compared with the reference fin profile geometry.

2 Numerical model development

2.1 Heat exchanger design

The heat exchanger used in the present study is fin and tube type. The design entails circular tubes and rectangular fins which are attached to the set of two tubes with a fixed gap in between. This particular design is also called 'H-type' finned tube heat exchanger due to the typical arrangement of fins on tubes resembling the letter 'H'. An orderly arrangement of the single units

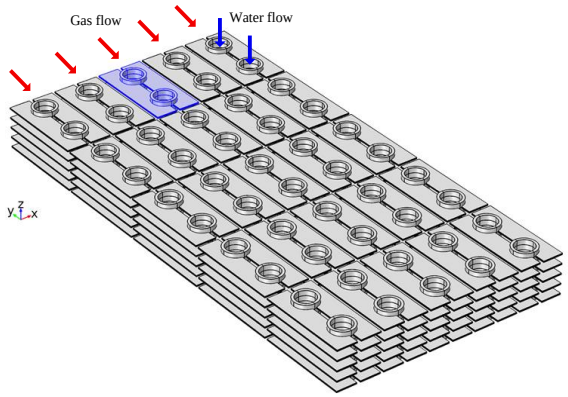


Figure 1. Double fin and tube (or H-type) heat exchanger configuration.

results in the complete heat exchanger configuration which can be scaled for desired applications based on the heat transfer rate and allowable pressure loss. Figure 1 shows the pictorial view of fin and tube heat exchanger configuration used in the present study. The design typically used in waste heat recovery applications such as marine boilers, where hot exhaust gas flows over the finned tube bundle, and cold water flows inside the tubes as can be seen in Figure 1. The heat transfers from hot exhaust gases, by convection through fins and conduction within fin and tube thickness, to the water inside the tubes for steam generation for other application purposes

2.2 Computational geometry

The geometry of the fin and tube heat exchanger simulated in the present study is shown in Figure 2. In order to save the computational effort, the geometry to be studied is reduced to one-half of the single unit. The computational geometry is divided into three domains- fin, tube and gas; and boundaries- inlet, outlet, and symmetry. The geometric dimensions of the heat exchanger design are given in Table 1.

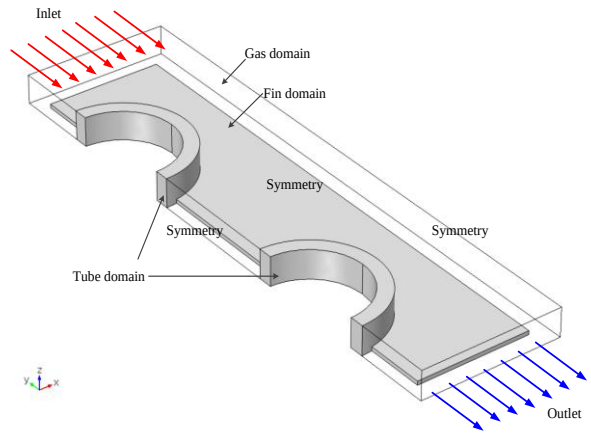


Figure 2. Computational geometry used in the present investigation.

Table 1. Design parameters and operating conditions for a single unit of the exchanger.

Parameter	*Symbol*	*Value*	*Unit*
Length of the fin	L_f	0.145	m
Width of the fin	W_f	0.070	m
Thickness of the reference fin base	$\delta_{fb,r}$	0.002	m
Thickness of the reference fin tip	$\delta_{ft,r}$	0.002	m
Width of the gap between fins	δ_a	0.007	m
Inner diameter of the tube	D_i	0.030	m
Outer diameter of the tube	D_o	0.038	m
Tube pitch	p_t	0.077	m
Length of the gas domain	L_g	0.155	m
Width of the gas domain	W_g	0.080	m
Fin pitch	p_f	0.015	m
Temperature at gas inlet	T_{in}	573.15	K
Pressure at gas outlet	p_{out}	0.0	Pa
Temperature of inner tube wall	T_w	453.15	K

2.3 Formulation of the fin geometric profile

In the present work, the geometry of the fin is varied using aspect ratio (α) as a profile parameter which is defined as the ratio of thickness of fin tip (δ_{ft}) to the thickness of fin base (δ_{fb}) and can be expressed as-

$$\alpha = \frac{\delta_{ft}}{\delta_{fb,r}} \quad (1)$$

The rectangular geometry of the fin is considered as a reference profile to simplify the analysis and geometric complexity, and the thickness of fin base is kept constant as of reference rectangular fin ($\delta_{fb,r}$) while the thickness of the fin tip is subjected to a variation (Figure 3). The aspect ratio is varied in a range α=1.0-0.1 transforming the reference rectangular fin profile (at α=1.0) into the trapezoidal profile (at α= 0.7, 0.5, 0.3) which eventually resembles a triangular profile (at α=0.1). With the change in aspect ratio, total heat transfer area, the thermal contact area between the fin and tubes and, the weight of the heat exchanger unit (computational geometry) changes as shown in Figure 4.

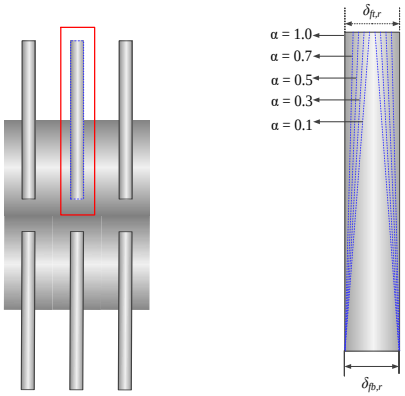


Figure 3. Computational geometry used in the present investigation.

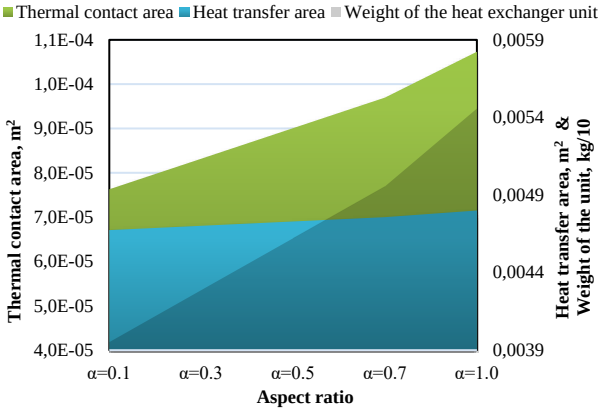


Figure 4. Computational geometry used in the present investigation.

2.4 Governing equations

3D CFD model is developed using commercially available Multiphysics software COMSOL v5.2. Following assumptions are made in the present model-

- Steady state flow and heat transfer
- Incompressible flow
- Negligible thermal contact resistance
- Temperature dependent fluid property
- Constant inner tube wall temperature
- No periodic boundary condition (i.e. model is valid for the first unit of the heat exchanger as shown in Figure1).

The mass and momentum balance for flow in the gas domain and energy balance in terms of heat transfer are given as-

$$\nabla \cdot \mathbf{u} = 0 \qquad (2)$$

$$\rho(\mathbf{u}\cdot\nabla)\mathbf{u} = \nabla\cdot[-p\mathbf{I} + \mu(\nabla\mathbf{u} + (\nabla\mathbf{u})^T)] + \mathbf{F} \qquad (3)$$

$$\rho C_p \mathbf{u}\cdot\nabla T + \nabla\cdot\mathbf{q} = \mathbf{Q} \qquad (4)$$

where, $\mathbf{q} = -k\nabla T$

Based on the mass flow rate and the heat exchanger configuration, the Shear Stress Transport (SST) model

Table 2. Design parameters and operating conditions for a single unit of the exchanger.

Initial	*Condition*
Gas domain	$\mathbf{u}=0\,;\,p=0\,;k=\left(\frac{10\cdot\mu}{\rho(0.1\cdot l_{ref})}\right)^2\,;\,\varepsilon=\frac{C_\mu k_{init}^{3/2}}{0.1\cdot l_{ref}}\,;\,\omega_{init}=\frac{\sqrt{k_{init}}}{0.1\cdot l_{ref}}$
All domains	$T = 298.15\ K$
Boundary	*Condition*
Inlet	$T=T_{in};\ u=0, v=-u_{in}, w=0$
Wall	$\mathbf{u}\cdot\mathbf{n}=0;\nabla k_e\cdot\mathbf{n}=0\ ;\ \varepsilon=\frac{C_\mu^{3/4}k_e^{3/2}}{\kappa_v\delta_w}$
Inner tube wall	$T=T_w$
Outlet	$p=p_{out}, -\mathbf{n}\cdot\mathbf{q}=0\ ;\ \nabla k_e\cdot\mathbf{n}=0\ ;\ \nabla\varepsilon\cdot\mathbf{n}=0$
Symmetry	$\mathbf{u}\cdot\mathbf{n}=0\ ;\ -\mathbf{n}\cdot\mathbf{q}=0\ ;\ \nabla k_e\cdot\mathbf{n}=0\ ;\ \nabla\varepsilon\cdot\mathbf{n}=0$

is adopted. The governing equations of two-equation SST model are formulated in terms of k and ω as (Mentor, 1994)-

$$\rho\frac{\partial k}{\partial t} + \rho\mathbf{u}\cdot\nabla k = P - \rho\beta_o k\omega + \nabla\cdot((\mu+\sigma_k\mu_T)\nabla k) \qquad (5)$$

$$\rho\frac{\partial \omega}{\partial t} + \rho\mathbf{u}\cdot\nabla\omega = \frac{\rho\gamma}{\mu_T}P - \rho\beta\omega^2 + \nabla\cdot((\mu+\sigma_\omega\mu_T)\nabla\omega) + 2(1-f_{v1})\frac{\rho\sigma_{\omega 2}}{\omega}\nabla\omega\cdot\nabla k \qquad (6)$$

The default model parameters used to solve the governing equations are defined in the Appendix. Table 2 lists initial conditions for a steady state simulation (Mentor et al., 2003) and boundary conditions used to solve the computational model and achieving preliminary results numerically.

3 Model validation

3.1 Experimental validation

The validation of numerical results is performed using the experimentally developed correlations by Chen et al. (2014) and a comparison is shown in Figure 5.

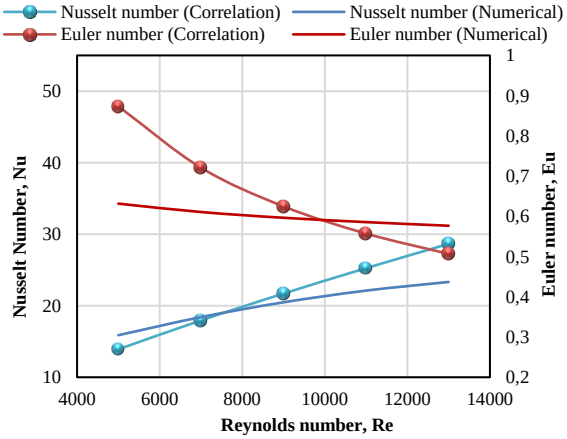


Figure 5. Computational geometry used in the present investigation.

Correlations for the Nusselt number and Euler number given by (7) and (8) are valid for Reynolds number range of 5000-18000 with a relative error of 2.79% and 3.70%, respectively. The average percent deviation of numerically predicted Nusselt and Euler numbers from the correlation values is calculated to be 5.61% and 5.72%, respectively. The deviation accounts for the assumptions in the present study or and the experimental errors in developing the correlations. These deviations are in acceptable range and hence, the results are assumed accurate enough to predict the physical behavior.

$$Nu = 0.053Re^{0.756}\left(\frac{D_o}{p_f}\right)^{-0.212}\left(\frac{L_f}{p_f}\right)^{-0.294}\left(\frac{W_f}{p_f}\right)^{0.155} \quad (7)$$

$$Eu = 19.14Re^{-0.57}\left(\frac{L_f}{D_o}\right)^{1.32} \quad (8)$$

3.2 Mesh independence test

Mesh independence test is made on the reference fin using temperature difference across the gas domain as an objective property. Five different meshes with 375860, 657449, 997272, 1716992 and 2130500 elements are used in the simulation. The test result suggests the mesh with 1716992 elements as a good choice in regards with the accuracy and computational time.

4 Results and discussion

The results predicted from the present study are discussed in this section. Table 3 expresses the dimensionless parameters used to evaluate the performance of the heat exchanger design. The Nusselt number and Euler number are used to assess the heat transfer and pressure loss characteristics of the heat exchanger with different fin geometric profile. As observed in Figure 6, the Nusselt number increases with the Reynolds number which shows thermal performance increases as the flow velocity increases. Moreover, the Nusselt number is higher for α=1.0, i.e., for rectangular

Table 3. Performance parameters.

Performance parameter	*Expression*
Nusselt number, *Nu*	$\frac{hD_o}{k}$
Euler number, *Eu*	$\frac{\Delta p}{\frac{1}{2}\rho_g u_{\max}^2}$
Efficiency index, η	$\frac{Nu}{\left(\frac{\Delta p}{\frac{1}{2}\rho_g u_{in}^2}\right)}$
Area-goodness factor, *j/f*	$\frac{\left(\frac{Nu}{RePr^{1/3}}\right)}{\left(\frac{\Delta p D_o}{\frac{1}{2}\rho_g u_{in}^2 L_g}\right)}$

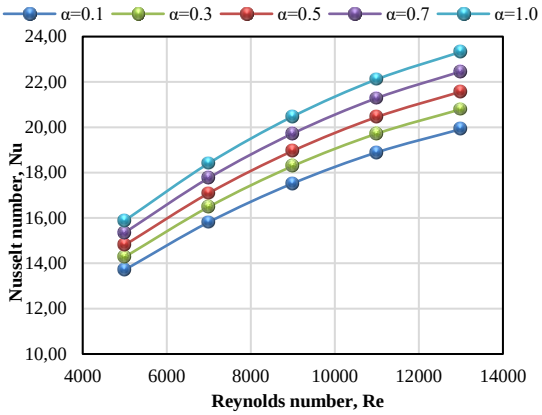


Figure 6. Variation of the Nusselt number with respect to Reynolds number.

fin profile and then decreases as α approaches to 0.1 where the fin geometric profile becomes nearly triangular. This effect results from the decreasing flow velocity as α varies from 1.0 to 0.1, which further decreases the convective heat transfer. Variation in Euler number with Reynolds number for different fin geometric profiles can be seen from Figure 7. Euler number decreases as α varies from 1.0 to 0.1, which is a clear demonstration of reduced pressure loss on a transition of rectangular fin profile (α=1.0) to triangular profile (α=0.1). At Re = 13000, Euler number for α=0.1 decreases by 7.23% than α=1.0.

In order to evaluate the overall performance of the heat exchanger in terms of both, heat transfer and pressure loss, efficiency index (Table III) is calculated. Figure 8 shows a variation in efficiency index with respect to Reynolds number for different fin geometric profiles. The efficiency index increases with the Reynolds number and so thus the overall performance of the heat exchanger design. As observed, efficiency index increases as α goes down from 1.0 to 0.1. which dictates that the fin with tapered geometric profile performs better in comparison to the conventional rectangular fin geometric profile.

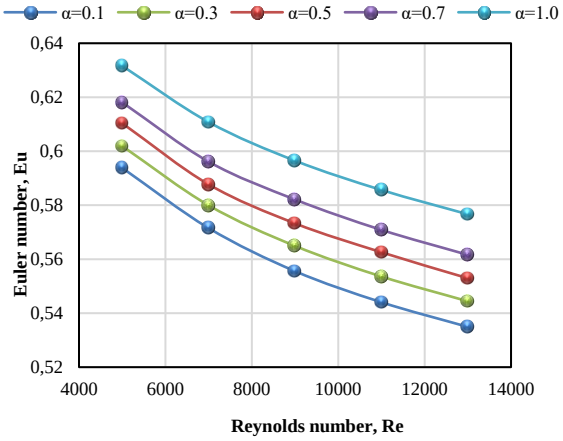


Figure 7. Variation of the Euler number with respect to Reynolds number.

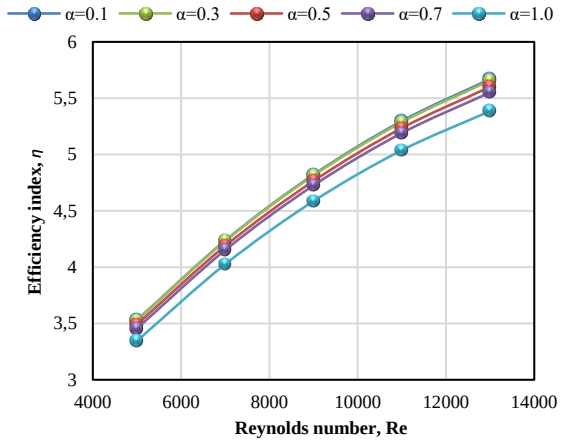


Figure 8. Variation of the efficiency number with respect to Reynolds number.

This trend of efficiency index explains that as α varies from 1.0 to 0.1, the pressure reduction is dominant than that of increment in thermal performance. In addition, the heat exchanger with α=0.1 and α=0.3 shows nearly equivalent performance. For instance, at Re=13000 efficiency index at α=0.1 is only 0.23 % higher than α=0.3. Heat transfer through the fin can be predicted from the temperature gradients on the fin surface.

Figure 9 shows temperature gradients on the fin surface of different geometric profiles. The dissipation of the heat from hot gas to the fin is evident from the higher temperatures away from fin and tube interface where the heat is conducted from the fin to the tube wall resulting in lower temperatures in those regions. Relatively, higher temperature gradients on the fin surface with α=0.1 are evident of lower heat transfer rate due to the lower temperature difference between the gas and the fin which further reduces the heat transfer performance.

To determine the impact of different fin profiles on the overall performance, a dimensionless parameter called 'area-goodness factor' is used. It is defined as a ratio of Colburn j factor to the friction factor, f of the heat exchanger design with respect to the reference fin geometry (Table III). Figure 10 shows the comparative performance of the heat exchanger with different fin geometric profiles at Reynolds number range 5000-13000. It can be observed that fin at α=0.1 (i.e., nearly triangular geometric profile) has the highest performance factor in comparison to the other fin profiles under similar operating conditions.

In many industrial applications of fin and tube heat exchangers such as waste heat recovery, aerospace, air-conditioning, automobile radiator, marine vessels, the available volume space and heat exchanger unit weight is a primary design consideration. Therefore, to investigate the most suitable geometric fin profile, the reduction in the weight of the heat exchanger unit as the fin geometric profile varies from α=1.0 to α=0.1 is determined and is shown in Figure 11. As α reduces from 1.0 the weight of the heat exchanger unit (considered one-half in the present study, see Figure 2) reduces and accounts for approximately 28 % reduction when α=0.1. Based on the results and discussion, it can be observed that fin geometric profile at α=0.1 when fin has nearly triangular geometric profile shows better performance with less weight than the reference rectangular fin geometry at α=1.0.

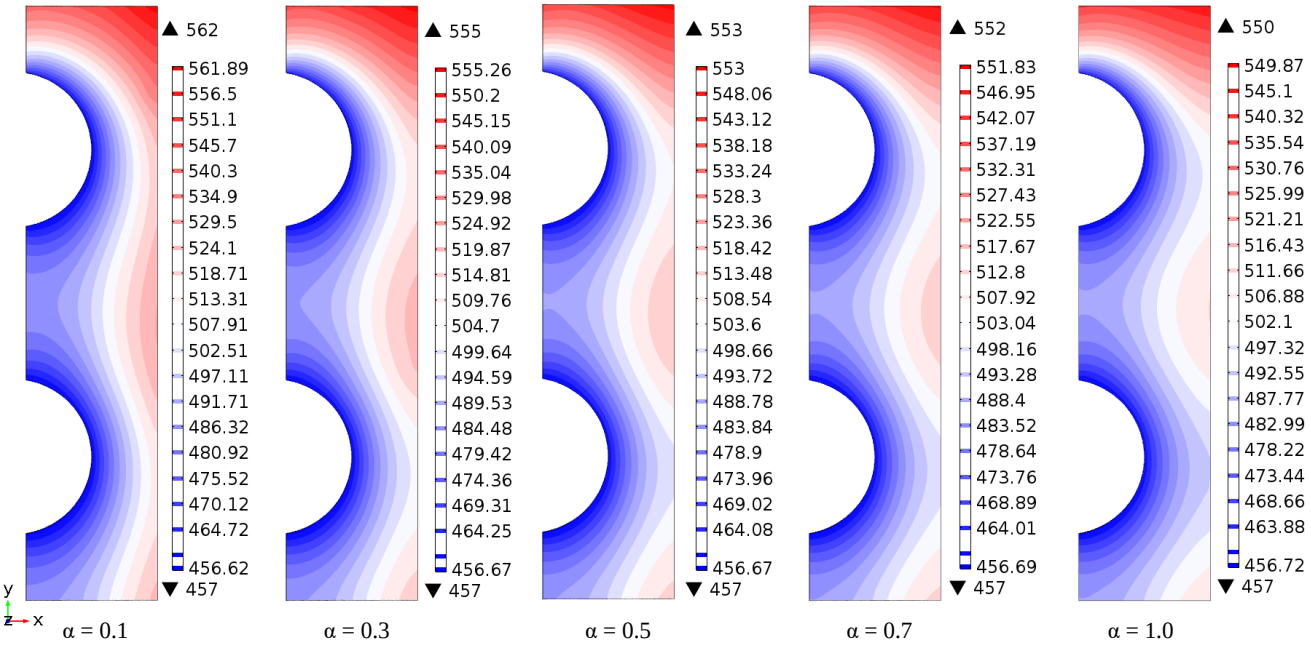


Figure 9. Temperature gradients on the fin surface of different geometric profiles.

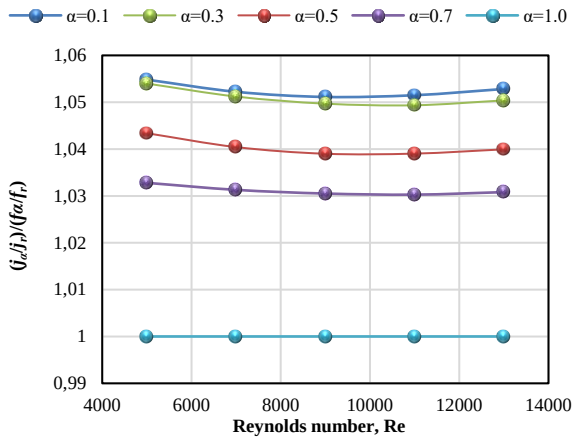


Figure 10. Performance comparison of different geometric fin profile.

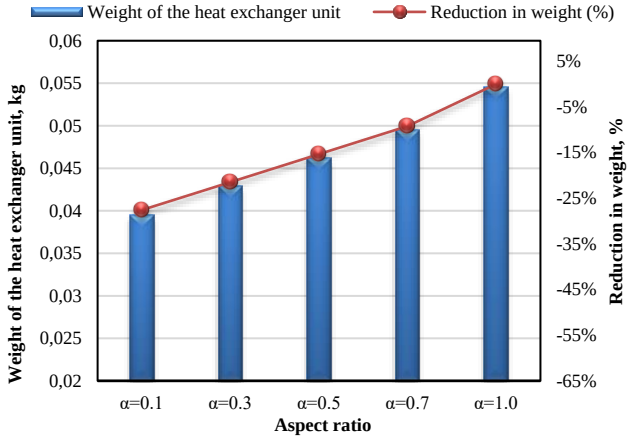


Figure 11. Comparison of the change in weight of heat exchanger unit with different geometric fin profile.

5 Conclusions

In the present study, the impact of different fin geometric profiles on the heat transfer performance and pressure loss in a fin and tube heat exchanger design are analyzed. The numerical study concludes that the fin with triangular at α=0.1 profile can enhance the heat transfer with reduced pressure loss in comparison to the conventional rectangular fin profile, α=1.0. Furthermore, the fin with α=0.1 reduces the heat exchanger weight up to 28 % which is always desirable in the industrial applications of fin and tube heat exchangers. The work presented in this paper encourages the further investigation of different possible fin geometric profiles to optimize the material and manufacturing cost which are the main controlling factors in designing a fin and tube heat exchangers at the industrial scale.

Acknowledgements

This work is a part of the research project: THERMCYC- Advanced thermodynamic cycles utilising low-temperature heat sources (Project No. 1305-0036B).

Nomenclature

Symbols

D	diameter of the tube, m
Eu	Euler number
$\mathbf{F}$	body force vector, N/m^3
h	Convective heat transfer coefficient, W/m^2.K
K	thermal conductivity, W/m.K
k	turbulent kinetic energy, m^2/s^2
L	length, m
Nu	Nusselt Number
p	pressure, Pa
Δp	pressure difference across the gas domain, Pa
Pr	Prandtl number
$\mathbf{Q}$	heat flux vector, W/m^2
Re	Reynolds number
T	temperature, K
Q	heat source or sink, W/m^3
u	flow velocity, m/s
$\mathbf{u}$	average velocity vector, m/s
ω	Specific dissipation rate, 1/s
ρ	density, kg/m^3
μ	dynamic viscosity of the gas, Pa.s

Subscripts

g	gas or gas domain
l	liquid
f	fin
w	inner tube wall
i	inner tube
o	outer tube
max	maximum
r	reference fin geometric profile

References

Y. A. Cengel, J. M. Cimbala, and R. H. Turner. *Fundamentals of thermal-fluid sciences*, Fourth edition in SI units. *McGraw-Hill*. 2012.

H. T. Chen, and J. R. Lai. Study of heat-transfer characteristics on the fin of two-row plate finned-tube heat exchangers. *Int. J. Heat Mass Tran.*, 55: 4088–4095, 2012. doi:10.1016/j.ijheatmasstransfer.2012.03.050.

H. Chen, Y. Wang, Q. Zhao, H. Ma, Y. Li, and Z. Chen. Experimental Investigation of Heat Transfer and Pressure Drop Characteristics of H-type Finned Tube Banks. *Energies*, 7(11): 7094-7104, 2014. doi:10.3390/en 7117094.

Y. Jin, G. H. Tang, Y. L. He, and W. Q. Tao. Parametric study and field synergy principle analysis of H-type finned tube bank with 10 rows. *Int. J. Heat Mass. Tran.*, 60: 241–251, 2013. doi: 10.1016/j.ijheatmasstransfer.2012.11.043.

F. R. Menter. Two-Equation Eddy-Viscosity Turbulence Models for Engineering Applications. *AIAA Journal*, 38(8): 1598-1605, 1994. doi:10.2514/3.12149.

F. R. Menter, M. Kuntz, and R. Langtry. Ten years of industrial experience with the SST Turbulence Model. *Turbulence Heat and Mass Transfer*, 2003. doi: 10.1.1.460.2814.

L. Tong. 3D numerical analysis of heat transfer characteristics for H-type finned tube. *Mech. Electr. Eng. Mag.*, 152: 79–81, 2007.

Y. C. Wang, and G. H. Tang. Acid condensation and heat transfer characteristics on H-type fin surface with bleeding

dimples and longitudinal vortex generators. *Chin. Sci. Bull.*, 59(33): 4405–4417, 2014. doi:10.1007/s11434-014-0564-3.

X. Yu, Y. Yuan, Y. Ma, and H. Liu. Experimental tests and numerical simulation on heat transfer and resistance characteristics of H-type finned tube banks. *J. Power Eng*., 30: 433–438, 2010. doi:10.1016/j.egypro.2017.03.1014.

Z. Zhang, Y. Wang, and Q. Zhao. Numerical study on performance optimization of H-type finned tubes. *J. Power Eng*., 30: 941–946, 2010.

X. B. Zhao, G. H. Tang, X. W. Ma, Y. Jin, and W. Q. Tao. Numerical investigation of heat transfer and erosion characteristics for H-type finned oval tube with longitudinal vortex generators and dimples. *Appl. Energy*, 127: 93–104, 2014. doi:10.1016/j.apenergy.2014.04.033.

Appendix

In (5), $P = \min(P_k, 10\rho\beta_o k\omega)$ (i)

where, $P_k = \mu_T(\nabla\mathbf{u} : (\nabla\mathbf{u} + (\nabla\mathbf{u})^T) - \frac{2}{3}(\nabla\cdot\mathbf{u})^2 - \frac{2}{3}\rho k\nabla\cdot\mathbf{u}$ (ii)

The turbulent viscosity is, $\mu_T = \frac{\rho a_1 k}{\max(a_1\omega, Sf_{v2})}$ (iii)

where, S is the characteristic magnitude of the mean velocity gradients,

$S = \sqrt{2S_{ij}S_{ij}}$

The other model constants are given in terms of interpolation functions as,

$\phi = f_{v1}\phi_1 + (1 - f_{v1})\phi_2$ for $\phi = \beta, \gamma, \sigma_k, \sigma_\omega$ and $f_{v1} = \tanh(\theta_1^4)$

$$\theta_1 = \min\left[\max\left(\frac{\sqrt{k}}{\beta_o\omega l_\omega}, \frac{500\mu}{\rho\omega l_\omega^2}\right), \frac{4\rho\sigma_{\omega 2}k}{CD_{k\omega}l_\omega^2}\right] \quad \text{(iv)}$$

where, lw is the distance to the closest wall.

$$CD_{k\omega} = \max\left(\frac{2\rho\sigma_{\omega 2}}{\omega}\nabla\omega\cdot\nabla k, 10^{-10}\right) \quad \text{(v)}$$

$$f_{v2} = \tanh(\theta_2^2) \text{ and } \theta_2 = \max\left(\frac{2\sqrt{k}}{\beta_o\omega l_\omega}, \frac{500\mu}{\rho\omega l_\omega^2}\right) \quad \text{(vi)}$$

The other default model parameter values are,

$\beta_1 = 0.075, \gamma_1 = 5/9, \sigma_{k1} = 0.85, \sigma_{\omega 1} = 0.5,$

$\beta_2 = 0.0828, \gamma_2 = 0.44, \sigma_{k2} = 1.0, \sigma_{\omega 2} = 0.856,$

$\beta_o = 0.09, a_1 = 0.31$ (vii)

The Reynolds number is calculated as: (viii)

$$Re = \frac{\rho u_{\max} D_o}{\mu}$$

The gas-side convective heat transfer coefficient is determined by overall heat transfer coefficient as:

$$\frac{1}{U} = \frac{1}{h_g} + \frac{A_t}{A_i}\left(\frac{1}{h_i} + \frac{(D_o - D_i)}{2K}\right) \quad \text{(ix)}$$

On further simplification,

$$\frac{1}{UA_t} = \frac{1}{h_g A_t} + \frac{1}{h_i A_i} + \frac{1}{\left[\frac{2KA_t}{(D_o - D_i)}\right]} \quad \text{(x)}$$

Since the heat transfer coefficient inside the tube is high (~10^4 W/m^2.K), the second term in Eq. (x) is omitted. The above equation can be further simplified without losing accuracy as the tubes being analyzed are of small thickness (~10^{-3} m) and higher thermal conductivity (~50 W/m.K) which makes the third term very small and hence negligible. This results in a much more straightforward expression,

$$\frac{1}{UA_t} = \frac{1}{h_g A_t} \quad \text{(xi)}$$

$$U = h_g \quad \text{(xii)}$$

Overall heat transfer coefficient can be determined as:

$$U = \frac{Q_t}{A_t \Delta T_{lm}} \quad \text{(xiii)}$$

$$\text{where, } \Delta T_{lm} = \frac{(T_w - \bar{T}_{in}) - (T_w - \bar{T}_{out})}{\ln\left[\frac{(T_w - \bar{T}_{in})}{(T_w - \bar{T}_{out})}\right]} \quad \text{(xiv)}$$

Voltage Stability Assessment of the Polish Power Transmission System

Robert Lis

Faculty of Electrical Engineering, Wroclaw University of Science and Technology, Poland,
Robert.Lis@pwr.edu.pl

Abstract

PSE S.A. is the sole Transmission System Operator in Poland and, as such, responsible for the provision of reactive power resources for maintaining the voltage within predefined limits. This paper describes the problems associated with the investigation of voltage stability of transmission power grid. Voltage problems are the result of heavy loading of transmission lines and transformers. Voltage instability has been responsible for voltage damage in some parts of Polish Power Transmission System (PPTS) on 26 June 2006. The voltage criteria used for voltage security assessment should require, that the worst bus voltage at post-contingency N-1 and sometimes N-2, must be approximately greater than 0.95 p.u. for generator buses and 0.9 p.u. for others. At the stage of planning, the active power transfer margin may be used as a proximity measure of voltage collapse.

Keywords: power system control, reactive power control, load flow control, voltage stability

1 Introduction

The idea of P-V and Q-V curve is used to determine the maximal reactive margin at load buses to avoid voltage instability. Sometimes the voltage stability study may be limited to identify the violation of the bus voltage constraints. In this paper the p-q curve for the critical bus voltage magnitude is created. Using this p-q curve the probability of the critical voltage violation is estimated for uniformly distributed active and reactive power at a given load bus. The p-q curve is created on the basis of bus impedance, which can be measured or calculated. To illustrate the usefulness of p-q idea the simple numerical example is presented. The paper describes also the importance of reactive power control basing on the failures and control problems in the PPTS during a dry summer period.

Assessing and mitigating problems associated with voltage security remains a critical concern for many power system planners and operators. Since it is well understood that voltage security is driven by the balance of reactive power in a system, it is of particular interest to find out what areas in a system may suffer reactive power deficiencies under some conditions and to obtaining information regarding how system voltage stability can be improved most effectively. Operation near the voltage stability limits is impractical and a sufficient power or voltage margin is needed. Practically, the idea of P-V and Q-V curve is used to determine the minimal margin to avoid voltage collapse (Chayapathi et al., 2013; Khoi et al., 1999; Lis, 2013).

Voltage stability is concerned with the ability of a power system to maintain acceptable voltages at all buses in the system under normal conditions and after being subjected to a disturbance (Taylor, 1994). As an example, Table 1 shows the voltage limits, which should be fulfilled in PPTS (Lis, 2013). According to their idea, the Thevenin's impedance is equal to the bus load impedance at the point of voltage collapse. In this paper the idea of using Thevenin's impedance to bus voltage study is extended by taking into account the bus load.

Table 1. Voltage Criteria in PPTS.

No.	*Bus*	*Normal Conditions*	*N-1 and N-2 Contingencies*
1	Generation buses 110 kV	1.0000p.u. - 1.1100 p.u.	0.9545 p.u. - 1.1100 p.u.
2	Generation buses 220 kV	1.0000 p.u. - 1.1136 p.u.	0.9545 p.u. - 1.1136 p.u
3	Generation buses 400 kV	1.0000 p.u. - 1.0500 p.u.	0.9500 p.u. - 1.0500 p.u.
4	Load buses 110 kV	0.9545 p.u. - 1.1100 p.u.	0.9000 p.u. - 1.1100 p.u.
5	Load buses 220 kV	0.9545 p.u. - 1.1136 p.u.	0.9091 p.u. - 1.1136 p.u.
6	Load buses 400 kV	0.9500 p.u. - 1.0500 p.u.	0.9000 p.u. - 1.0500 p.u.

From the point of view of monitoring and control the following transmission constraints are the most important: thermal limit of lines of transmission subsystem, voltage stability limit of transmission subsystem and angle stability limit of transmission subsystem. Voltage problems occur in heavily stressed

power systems. Then, the voltage stability limit may be sometimes more drastically important than thermal limits. This is the case of the voltage collapse in PPTS on June 26, 2006.

This paper is devoted to the analysis of voltage limits. The main question is how far we are from the voltage instability and how to consider the randomness of loads. The original p-q curve is applied here to solve such a task. The idea of using p-q curve for voltage collapse analysis was presented in this paper. The p-q curve for the critical bus voltage magnitude is created. Using this p-q curve the critical voltage violation is estimated for uniformly distributed active and reactive power at a given load bus.

The p-q curve is created on the basis of bus impedance. The mathematical background of the proposed idea is presented. To illustrate the usefulness of p-q idea the simple numerical example is presented.

1.1 Country-Wide Absence of Electrical Supply– a Blackout

The quality of the electrical energy supply can be evaluated basing on a number of parameters (Abril et al., 2003; Yorino et al., 2003; Shubhanga et al., 2002). However, the most important will be always the presence of electrical energy and the number and duration of interrupts. If there is no voltage in the socket nobody will care about harmonics, sags or surges. A long term, wide-spread interrupt - a blackout leads usually to catastrophic losses. It is difficult to imagine that in all the country there is no electrical supply. In reality such things have already happened a number of times.

One of the reason leading to a blackout is reactive power, that went out of the control. When consumption of electrical energy is high, the demand on inductive reactive power increases usually at the same proportion. In this moment, the transmission lines (that are well loaded) introduce an extra inductive reactive power. The local sources of capacitive reactive power become insufficient. It is necessary to deliver more of the reactive power from generators in power plants. It might happen that they are already fully loaded and the reactive power will have to be delivered from more distant places or from abroad. Transmission of reactive power will load more the lines, which in turn will introduce more reactive power. The voltage on customer side will decrease further. Local control of voltage by means of autotransformers will lead to increase of current (to get the same power) and this in turn will increase voltage drops in lines. In one moment this process can go like avalanche reducing voltage to zero. In mean time most of the generators in power plants will switch off due to unacceptably low voltage what of course will deteriorate the situation.

In continental Europe, most of the power plant are based on heat and steam turbines. If a generation unit in such power plant is stopped and cool down it requires time and electrical energy to start operation again. If the other power plants are also off - the blackout is permanent (Bhattacharya et al., 2001, 2002; Chicco et al., 2013).

The difficulties showed up on summer 2006. The prediction for power consumption on this day was 18200 MW (in the morning peak) what was much higher compared with June in last year or previous years. This power was planned to be supplied from 75 generation units. Above these, there were a hot power reserve of 1350 MW (in this 237 MW second-reserve, 656 MW minute-reserve) and a cold reserve of about 2600 MW. In the north-east Poland there is not any grid-generation. The closest to this region is Ostroleka Power Plant (P. P.), which in that time from three 200 MW units has two in operation and one set off for maintenance. In early morning of the Jun 26th one unit in Power Plant Patnow had to be switched off and before noon four other units (two in Kozienice P. P. and two in Laziska P. P.) were switched off as well. All these unites were the main supplier to the north-east region of Poland. At 7 o'clock 570 MW of power was lost. At the same time the consumption prediction appeared to be wrong - the consumption was 600 MW higher and there was also much higher demand on reactive power. At 13 o'clock there was an unbalance of 1100 MW. In mean time one unit (in Dolna Odra P. P.) had been activated. However further activation from cold-reserve required more time (about 6 hours) because of technological reasons.

Unusual heat wave spreading throughout the country caused deterioration of the operational conditions in power plants. Due to lack of sufficient amount of cooling water and exceeded water temperature levels, the generating capacities of some power plants systematically decreased. That situation concerned mainly the power plants located in the central and northern part of Poland, the loadings of some transmission lines reached the acceptable limits what in turn cause the necessity of generation decrease in power plants located outside the mentioned region. The control of reactive power became critical (Hatziadoniu et al., 2003; Lu et al., 2002).

2 Prepare Bus Load Flow Equations Using Thevenin's Circuit

Thevenin's theorem states that in the linear electric circuit the effect of the load change at a given bus can be represented by a simple circuit with *emf* $\underline{E}_T$ and the bus impedance $\underline{Z}_T$. The basic circuit resulting from Thevenin's theorem is shown in Figure 1. Knowing the Thevenin's bus impedance $\underline{Z}_T = R_T + jX_T$, load bus voltage V, active P and reactive bus power Q one can calculate the magnitude of Thevenin's *emf* using the following formula:

$$E_T = \sqrt{\left(V + \frac{PR_T + QX_T}{V}\right)^2 + \left(\frac{PX_T - QR_T}{V}\right)^2} \quad (1)$$

Using complex notification:

$$\underline{V} = V_a + jV_b \quad and \quad V = \sqrt{V_a^2 + V_b^2} \quad (2)$$

$$\underline{E}_T = E_a + jE_b \quad and \quad E_T = \sqrt{E_{Ta}^2 + E_{Tb}^2} \quad (3)$$

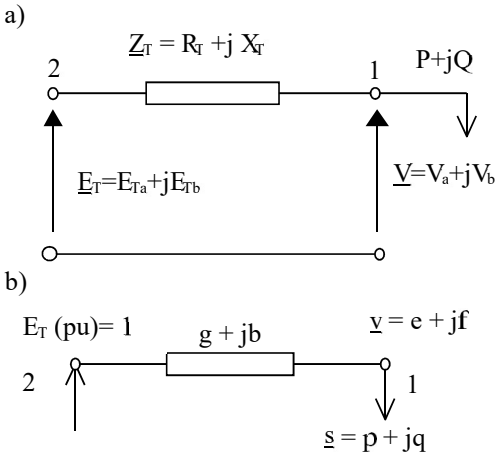


Figure 1. The scheme of 2-bus Thevenin's network, a) simple Thevenin's circuit, b) scheme for load flow study.

The complex admittance of branch connecting load bus with Thevenin's *emf* bus equals:

$$G_T + jB_T = 1/(R_T + jX_T) \quad (4)$$

The load flow equations for load bus have the following form:

$$P = V^2 G_{11} + (V_a E_a + V_b E_b) G_{12} + (-V_a E_b + V_b E_a) B_{12} \quad (5)$$

$$Q = -V^2 B_{11} - (V_a E_a + V_b E_b) B_{12} + (-V_a E_b + V_b E_a) G_{12} \quad (6)$$

where $$G_{11} = G_T \quad and \quad G_{12} = -G_T \quad (7)$$

$$B_{11} = B_T \quad and \quad B_{12} = -B_T \quad (8)$$

Let the bus 2 with Thenenin's *emf* be the slack bus. Then I have

$$E_a = E_T \quad and \quad E_b = 0 \quad (9)$$

and $$P = V^2 G_{11} + V_a G_{12} + V_b B_{12} \quad (10)$$

$$Q = -V^2 B_{11} - V_a B_{12} + V_b G_{12} \quad (11)$$

or $$P = V^2 G_T - V_a G_T - V_b B_T \quad (12)$$

$$Q = -V^2 B_T + V_a B_T - V_b G_T \quad (13)$$

To simplify all considerations the load bus per unit system is introduced as follows

$$Z_b = Z_T = \sqrt{R_T^2 + X_T^2} \quad (14)$$

$$V_{base} = E_T \quad (15)$$

$$S_b = E_T^2 / Z_b \quad (16)$$

where symbol *b* means the base value. Dividing both side of load flow equations by S_b I obtain:

$$p = v^2 g - eg - fb \quad (17)$$

$$q = -v^2 b + eb - fg \quad (18)$$

where $$p = P / S_b \quad and \quad q = Q / S_b \quad (19)$$

$$v = V / E_T \quad (20)$$

$$e = V_a / E_T \quad and \quad f = V_b / E_T \quad (21)$$

$$g = G_T Z_T \quad and \quad b = B_T Z_T \quad (22)$$

Note that the following relations exist in the new bus load per unit system:

$$r + jx = R_T / Z_T + jX_T / Z_T \quad (23)$$

$$z^2 = r^2 + x^2 = \frac{R_T^2}{Z_T^2} + \frac{X_T^2}{Z_T^2} = 1 \quad (24)$$

$$\underline{y} = 1/\underline{z} = 1/(r + jx) = (r - jx)/z^2 = r - jx \quad (25)$$

$$\underline{y} = g + jb \quad (26)$$

And finally $$g = r \quad and \quad b = -x \quad (27)$$

Hence the load flow equations in load bus per unit are as follows:

$$p = v^2 r - er + fx \quad (28)$$

$$q = v^2 x - ex - fr \quad (29)$$

2.1 The P-Q Curve for the Critical Bus Voltage Magnitude

From the point of view of voltage stability the voltage magnitude at a given bus must be in the range of upper and lower voltage limit

$$V_{lower} \le V \le V_{upper} \tag{30}$$

Especially the lower value is the critical value V_{cr} from the point of view of avoiding voltage instability in the power system. Hence the bus voltage must be greater then the critical value

$$V > V_{cr} \tag{31}$$

Using the new per unit system I can write

$$v > v_{cr} \tag{32}$$

where $$v_{cr} = V_{cr} / E_T \tag{33}$$

According to the above assumptions load flow equations for load bus critical solutions (e,f) depend on the critical voltage magnitude vcr. The load flow equations can be analyzed as a critical p-q curve composed of (p,q) values, which are related to the critical bus voltage magnitude vcr .

To find the formula of the critical p-q curve I must eliminate the rectangular components of *e* and *f* from the load flow equations (28) and (29). To find *e* I can make the following multiplications

$$rp = r^2 v^2 - r^2 e + rxf \tag{34}$$

$$xq = x^2 v^2 - x^2 e - rxf \tag{35}$$

and the following addition

$$rp + xq = z^2 v^2 - z^2 e \tag{36}$$

For $z = 1$ I have finally

$$e = v^2 - (rp + xq) \tag{37}$$

To find *f* I can made the following multiplications

$$xp = rxv^2 - rxe + x^2 f \tag{38}$$

$$rq = rxv^2 - rxe - r^2 f \tag{39}$$

and the following subtraction

$$xp - rq = z^2 f \tag{40}$$

For $z = 1$ I have finally

$$f = xp - rq \tag{41}$$

Substituting the obtained formula of *e* and *f* to the formula of v_{cr} I have as follows

$$v_{cr}^2 = e^2 + f^2 \tag{42}$$

$$v_{cr}^2 = (v_{cr}^2 - (rp + xq))^2 + (xp - rq)^2 \tag{43}$$

$$\begin{aligned} v_{cr}^2 &= v_{cr}^4 - 2v_{cr}^2 (rp + xq) \\ &\quad + (rp + xq)^2 + (xp - rq)^2 \end{aligned} \tag{44}$$

$$\begin{aligned} v_{cr}^2 &= v_{cr}^4 - 2v_{cr}^2 rp - 2v_{cr}^2 xq \\ &\quad + r^2 p^2 + 2rxpq + x^2 q^2 \\ &\quad + x^2 p^2 - 2rxpq + r^2 q^2 \end{aligned} \tag{45}$$

$$v_{cr}^2 = v_{cr}^4 - 2v_{cr}^2 rp - 2v_{cr}^2 xq + p^2 + q^2 \tag{46}$$

Hence, I obtain the following formula of p-q curve for a critical voltage magnitude v_{cr}

$$p^2 + q^2 - 2v_{cr}^2 rp - 2v_{cr}^2 xq + v_{cr}^4 - v_{cr}^2 = 0 \tag{47}$$

From the above formula I can obtain the quadratic equations for the specific *p*

$$q^2 - 2v_{cr}^2 xq + p^2 - 2v_{cr}^2 rp + v_{cr}^4 - v_{cr}^2 = 0 \tag{48}$$

An example of p-q curve is shown in Figure 2. The p-q curve can be transformed into P-Q curve after multiplication *p* and *q* by S_b .

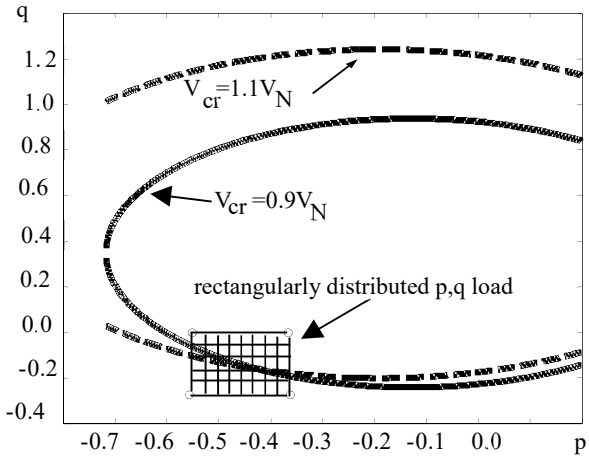


Figure 2. The example p-q curve at a load bus. Symbol V_N means the nominal voltage.

Equation (48) is quadratic and has two real solutions according to the value of the equation discriminant Δ:

$$\Delta = 4v_{cr}^4 x^2 - 4p^2 + 8v_{cr}^2 rp - 4v_{cr}^4 + 4v_{cr}^2 \tag{49}$$

$$\Delta = 4v_{cr}^4 x^2 - 4p^2 + 8v_{cr}^2 rp - 4z^2 v_{cr}^4 + 4v_{cr}^2 \quad (50)$$

$$\Delta = 4v_{cr}^4 x^2 - 4p^2 + 8v_{cr}^2 rp - 4r^2 v_{cr}^4 - 4x^2 v_{cr}^4 + 4v_{cr}^2 \quad (51)$$

$$\Delta = -4p^2 + 8v_{cr}^2 rp - 4r^2 v_{cr}^4 + 4v_{cr}^2 \quad (52)$$

$$\Delta = 4v_{cr}^2 - 4(p^2 - 2v_{cr}^2 rp + r^2 v_{cr}^4) \quad (53)$$

$$\Delta = 4v_{cr}^2 - 4(p - v_{cr}^2 r)^2 \quad (54)$$

$$\sqrt{\Delta} = 2\sqrt{v_{cr}^2 - (p - rv_{cr}^2)^2} \quad (55)$$

and finally I obtain two parts of p-q curve

$$q_{1,2} = 0.5\left(2v_{cr}^2 x - / + \sqrt{\Delta}\right). \quad (56)$$

The lower part of the p-q curve is associated with the consumed power, because a reactive consumed power at bus is treated in load flow equations as a negative value

$$q_{lower} = v_{cr}^2 x - \sqrt{-p^2 + 2v_{cr}^2 rp + v_{cr}^2 - v_{cr}^4 r^2} \quad (57)$$

The upper part of the p-q curve relates to positive values, i.e. to reactive generation at a given bus

$$q_{upper} = v_{cr}^2 x + \sqrt{-p^2 + 2v_{cr}^2 rp + v_{cr}^2 - v_{cr}^4 r^2} \quad (58)$$

2.2 The Probability of the Violation of the Critical Voltage

Let's assume that the load at bus is uniformly distributed between their min and max

$$p_{min} \le p \le p_{max} \quad (59)$$

$$q_{min} \le q \le q_{max} \quad (60)$$

The probability of the violation of the critical voltage magnitude at load bus can be calculated using the outside area and the rectangular area, see Figure 3. To find the probability of the violation of the critical voltage the lower part of p-q curve should be used, Figure 3. Using the geometrical definition of probability:

$$p_{vcr} = S_{outside} / S \quad (61)$$

where S means the area of rectangular

$$S = (p_{max} - p_{min})(q_{max} - q_{min}) \quad (62)$$

The outside area equals

$$S_{outside} = S - S_{ABCD} \quad (63)$$

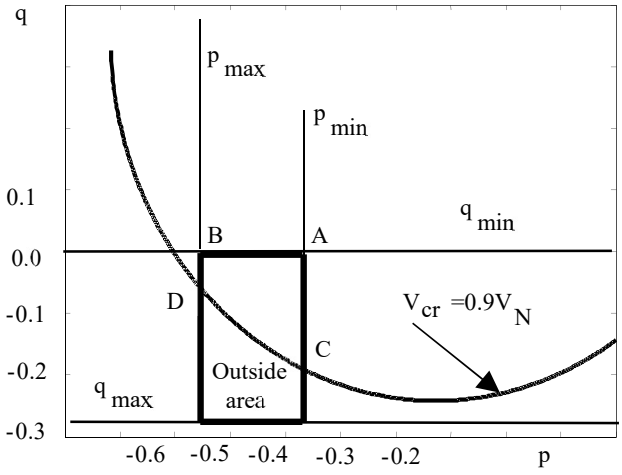


Figure 3. The p-q curve and the rectangularly distributed load.

The area of *ABC*D figure can be computed using the definite integral formula in the following way

$$S_{ABCD} = S1 - S2 = \int_{p\,max}^{p\,min} q_{min}\,dp - \int_{p\,max}^{p\,min} q_{lower}\,dp \quad (64)$$

where

$$S1 = \int_{p\,max}^{p\,min} q_{min}\,dp = q_{min}\,p\,\Big|_{p\,max}^{p\,min} = q_{min}(p_{min} - p_{max}) \quad (65)$$

and

$$\begin{aligned} S2 = {} & v_{cr}^2 x(p_{min} - p_{max}) + \\ & - \frac{(-p_{min} + v_{cr}^2 r)}{2}\sqrt{W_{min}} + \\ & + \frac{(-p_{max} + v_{cr}^2 r)}{2}\sqrt{W_{max}} + \\ & + \frac{v_{cr}^2}{2} arcsin\frac{(-p_{min} + v_{cr}^2 r)}{v_{cr}} + \\ & - \frac{v_{cr}^2}{2} arcsin\frac{(-p_{max} + v_{cr}^2 r)}{v_{cr}} \end{aligned} \quad (66)$$

Finally we can make the following substitution:

$$W_{min} = -p_{min}^2 + 2v_{cr}^2 rp_{min} + v_{cr}^2 - v_{cr}^4 r^2 \quad (67)$$

$$W_{max} = -p_{max}^2 + 2v_{cr}^2 rp_{max} + v_{cr}^2 - v_{cr}^4 r^2 \quad (68)$$

3 Numerical Example

Thevenin's complex impedance seen from the 400 kV load bus has been obtained by the load flow study in 400/220 kV transmission grid:

$$\underline{Z}_T = R_T + jX_T = (-125.12 + 31.40)\ \Omega.$$

The bus voltage magnitude at the analysed load bus determined by load flow computation equals: V = 410

kV, while the nominal voltage has the following value V_N = 400 kV. The critical voltage magnitude equals:

$$V_{cr} = 0.9V_N = 360 \text{ kV}.$$

The minimal and maximal active and reactive load at the analysed bus equal

$$P_{min} = 400 \text{ MW} \quad \text{and} \quad P_{max} = 600 \text{ MW};$$
$$Q_{min} = 0 \text{ MVAR} \quad \text{and} \quad P_{max} = 300 \text{ MVAR}.$$

Knowing the Thevenin's bus impedance $\underline{Z_T} = R_T + jX_T$, load bus voltage V, active P and reactive bus power Q I calculate the magnitude of Thevenin's *emf*:

$$E_T = \sqrt{\left(V + \frac{PR_T + QX_T}{V}\right)^2 + \left(\frac{PX_T - QR_T}{V}\right)^2} = 572.9\text{kV}.$$

To simplify all considerations the load bus per unit system is introduced: $Z_b = Z_T = \sqrt{R_T^2 + X_T^2} = 342.1\ \Omega$;

$$V_{base} = E_T = 572.9 \text{ kV};$$
$$S_b = E_T^2 / Z_b = 959.4 \text{ MVA}.$$

The value of analyzed variables in load by per unit systems are as follows:

$$p_{min} = P_{min} / S_b = \text{-0.4169};$$
$$p_{max} = P_{max} / S_b = \text{-0.6254};$$
$$q_{min} = Q_{min} / S_b = 0;\ q_{max} = Q_{max} / S_b = \text{-0.3127};$$
$$v_{cr} = V_{cr} / E_T = 0.6284;$$
$$r = R_T / Z_T = \text{-0.3657};\ x = X_T / Z_T = 0.9307.$$

Now can calculate the probability of the violation of the critical voltage magnitude. The rectangular area:

$$S = (p_{max} - p_{min})(q_{max} - q_{min}) = 0.0652.$$

The inside area: $S_{ABC} = \int_{p_{min}}^{p_{max}} (q_{lower} - q_{min})dp = 0.0270.$

The outside area $S_{outside} = S - S_{ABCD} = 0.0382.$

The probability of the violation of the critical voltage magnitude: $p_{vcr} = S_{outside} / S = 0.58$ □

4 Conclusions

The proposed p-q curve method is simple and may be based on local measurements of bus impedance. It enables calculating the probability of voltage limit violation at a given load bus. The greater the probability the weaker the bus is from the point of view of voltage stability.

To find the formula of p-q curve a new load bus per unit system must be introduced. The transformation from p-q curve to P-Q curve can be easily made by multiplication p and q value by the base power of the analyzed load bus.

The probability of voltage limit violation is estimated as the quotient of relevant area outside and inside the specific p-q curve.

The slower forms of voltage instability can be analyzed as steady state problem using power flow simulation. Snapshot in time following an outage may be simulated and P-U curves computed to assess voltage stability margin.

The main difficulties of the modelling of the power system concern load modelling and therefore conservative constant load hypothesis are used in computation.

Acknowledgements

This research was performed in cooperation with the Polish Power Grid Operator: PSE Operator S.A., Warszawska 165, 05-520 Konstancin-Jeziorna, Poland.

References

P. I. Abril, J. A. G. Quintero. VAR compensation by sequential quadratic programming. *IEEE Trans. on Power Systems*, 18:36-41, 2003. doi:10.1109/TPWRS.2002.807049.

K. Bhattacharya, J. Zhong. Reactive power as an ancillary service. *IEEE Trans. on Power Systems*, 16:294-300, 2001. doi:10.1109/59.918301.

K. Bhattacharya, J. Zhong. Reactive power management in deregulated electricity markets - a review. In conference proceedings: *IEEE Power Eng. Soc. Winter Meeting*, New York, 2002. doi: 10.1109/PESW.2002.985223.

V. Chayapathi, B. Sharath, G. S. Anitha. Voltage Collapse Mitigation by Reactive Power Compensation at the Load Side. *Ijret*, available via http://www.ijret.org, 02(09), 2013. doi: 10.15623/ijret.2013.0209037.

G. Chicco, G. Gross. Allocation of the reactive power support requirements in multitransaction networks. *IEEE Transactions on Power Systems*, 17:243-249, 2012.

C. J. Hatziadoniu, N. Nikolov, F. Pourboghrat. Power Conditioner Control and Protection for Distributed Generators and Storage. *IEEE Trans. on Power Systems*, 18:83-90, 2003.

Vu Khoi, M. M. Begovic, D. Novosel, M. M. Saha. Use of local measurements to estimate voltage stability margin. *IEEE Trans. on Power Systems*, 14:1029-1035, 1999.

R. Lis. *Problems of assessment and ways of improving voltage stability of an electrical power transmission grid*. Monograph, Wroclaw University of Technology Press. 2013. ISSN 0324-976x.

R. Lis. Voltage Stability Assessment Using Bus P-Q Curve. *Mathematics and Computers in Contemporary Science.* 2013. ISBN: 978-960-474-356-8

Y. Lu, A. Abur. Static security enhancement via optimal utilization of thyristor-controlled series capacitors. *IEEE Trans. on Power Systems*, 17:324-329, 2002.

K.N. Shubhanga, A. M. Kulkarni. Application of structure preserving energy margin sensitivity to determine the effectiveness of shunt and series FACTS devices. *IEEE Trans. on Power Systems*, 17:730-738, 2002.

C.W. Taylor, *Power System Voltage Stability*. McGraw-Hill. 1994.

N. Yorino, E. E. El-Araby. Sasaki and S. Harada. A new formulation for FACTS allocation for security enhancement against voltage collapse. *IEEE Trans. on Power Systems*, 18:3-10, 2003. doi: 10.1109/TPWRS.2002.804921.

Agglomeration Detection in Circulating Fluidized Bed Boilers using Refuse Derived Fuels

Nathan Zimmerman[1] Konstantinos Kyprianidis[1] Carl-Fredrik Lindberg[1,2]

[1]School of Business, Society and Engineering, Mälardalen University, Box 883, 72123 Västerås, Sweden, {nathan.zimmerman,konstantinos.kyprianidis}@mdh.se
[2]ABB Corporate Research, Forskargränd 8, 72178 Västerås, Sweden, carl-fredrik.lindberg@se.abb.com

Abstract

The formation of agglomerates in a refuse derived fuel (RDF) fired circulating fluidized bed (CFB) boiler has been investigated by implementing a dynamic model of the combustion process. The nature of refuse derived fuel, which is complex in composition, leads to an increased tendency for agglomerate formation. Notwithstanding the fact that a robust control scheme is essential in preventing the decrease in boiler efficiency from accelerated agglomerate formation. Therefore, a mechanism for detecting agglomeration through a physical model by looking at the minimum fluidization is presented. As agglomerates form between the fuel ash and bed sand the average diameter of the sand will increase and therefore the minimum fluidization velocity. Samples of bed material have been sieved and measured from a 160 MW circulating fluidized bed boiler fired with refuse derived fuel to determine bed material size distribution. The findings have been correlated and match an increase in the minimum fluidization velocity during a seven day sampling period where the bed material size distribution increases above the average sand diameter.

Keywords: circulating fluidized bed, agglomeration, detection, RDF

1 Introduction

The role of waste management in the 21st century is essential in maintaining the excessive amount of waste produced each year in the world. Most developed countries have modern routines and guidelines for dealing with municipal solid waste (MSW) by sorting out recyclable, toxic, and organic materials. Some of these countries are utilizing MSW for producing refuse derived fuels (RDFs). These are used in waste incinerators for the purpose of producing heat and power. When compared to other alternatives, such as land-filling, waste incineration is more reliable and environmentally friendlier (Hernandez-Atonal et al., 2007).

Due to the complex nature of RDF, a viable option for thermal waste treatment is to use them as a source of fuel in circulating fluidized bed (CFB) boilers to produce heat and power, simultaneously solving waste and energy issues. CFB boilers are unique in design and allow for a high degree in fuel flexibility, high combustion efficiency, lower emissions, and relatively fast response to load change (Basu and Fraser, 2015; Hernandez-Atonal et al., 2007; Gungor, 2009), and therefore can handle the complexities of RDF.

Typically, CFBs operate at a combustion temperature between 750 °C and 900 °C and use sand to fluidize the bed material. However, a large portion of the combustible material in RDF contain high levels of alkali and chlorine compounds, which both have low melting and low vaporization temperatures (Pettersson et al., 2013), and can react with the fluidizied sand to form low melting eutectics. It is crucial to maintain the operating temperature within limits. Else there is a significant risk of sand-ash reactions which can lead to agglomeration, slagging, and in the worse case scenario complete defluidization. The variability of RDF's composition, and therefore heating value, can lead to the temperature deviating above the boiler's limits for short periods, in the form of hot zones. The latter can increase the propensity of agglomerate formation either through ash melting or the ash coating of sand. Once agglomerates form they reduce the boiler's efficiency because the fluidized median is hindered and requires increasingly more air to properly combust the incoming fuel.

A dynamic model of an RDF fired CFB has been modeled after a 160 MW industrial installation. Bed samples for a one week period have been taken and analyzed for any increase in the bed material particle diameter. The model has proven to be capable of detecting agglomerate formation off-line, showing an increase in the minimum fluidization during a period of time when there was a corresponding significant increase in particle diameter.

2 Literature Review

2.1 Circulating Fluidized Bed Boilers

The schematic illustrated in Figure 1 is that of a CFB. Its a type of boiler that uses sand as the fluidizing median, where a degree of solid reflux is achieved allowing for a more uniform temperature throughout the boiler. Major advantages include stable combustion at low temperatures (750-900°C), uniform temperature distribution, large solid-gas exchange area, ability to handle problematic fuels with a large variability in size, moisture content,

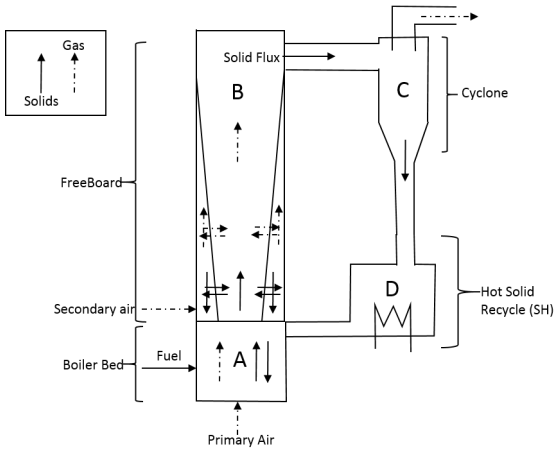


Figure 1. Circulating Fluidized Bed Boiler.

and heating value. They are capable of high heat transfer coefficients between bed and heat exchanger surfaces (Werther, 2007; Scala and Chirone, 2004).

When the fuel enters the boiler it begins to heat, where water is evaporated, volatiles begin to be released through devolatilization, ignition begins, and fragments begin to form. During this process more char becomes available for combustion. Primary air enters the bed of the boiler and entrains the sand and fuel. Above the bed is the freeboard, where secondary air is injected to allow for a higher combustion efficiency and it has been shown to reduce CO concentrations (Hernandez-Atonal et al., 2007). At the top of the boiler the flue gas along with a net solid flux of solids are carried to the cyclone which allows for the flue gas to be utilized in the convective section and separates fly ash from hot solids (sand and ash). The hot solids then pass through a heat exchanger to utilize their high enthalpy content, and then the cooled solids which contain unburnt char are recirculated back into the bed via the loop seal to be fully combusted.

2.2 Characteristics of RDF

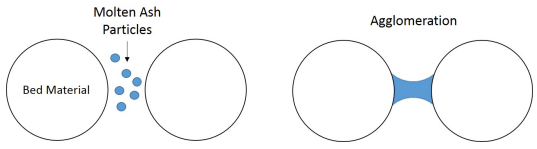
RDF is a byproduct of home and industrial waste, but before it can be fired it needs to be sorted and treated. The waste comes into the sorting facility were it is first chopped and shredded into credit card size pieces. Metals are removed by using high power magnets and the remaining mixture passes over a large wind-sifter, which allows for heavier objects such as glass and ceramics to dropout of the mixture. The separated items are sent to recycling, where what is left can be classified as RDF, and is ready for transport into the boiler.

We can characterize what remains in the fuel as either biomass-based (climate neutral) or fossil-based, where the fossil-based constituents are typically plastics and textiles. Due to the sorting process, determining the proportions of biomass vs. fossil based portions through manual sorting is impractical. There are three established methods for determining the biomass-based portion, and therefore the fossil based portions of RDF: The Selective Dissolution Method, The Balance Method, and the The ^{14}C-Method (Staber et al., 2008). The downside is that these methods are impractical for real-time applications.

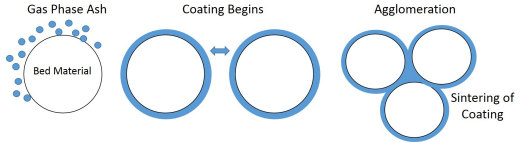
2.3 Agglomeration

The development of agglomerates, the binding of bed particles, is highly dependent upon the characteristics of the fuel's ash content, particle-to-particle interactions and the hydrodynamics in the boiler. As the bed material is fluidized ash can melt and bridge bed material together, as illustrated in Figure 2a. Alternatively, ash can lead to a buildup on bed material and create a sticky coating, as illustrated in Figure 2b, which causes more particles to stick together. Therefore, the likelihood of agglomeration formation is dependent upon ash characteristics and has been the topic in many studies looking at biomass and waste fuels (Elled et al., 2013; Lin et al., 2003; Scala and Chirone, 2008; Bajamundi et al., 2015; Ryabov et al., 2003). Proper fluidization is also needed to inhibit the likelihood of agglomeration. If particle mobility and mixing are below optimal it will not only cause a reduction in heat and mass transfer of the bed material (Liu et al., 2012), but the viscous materials can bond together and can lead to permanent bonds depending upon the residence time.

In a study by (Lindberg et al., 2013) they stated that the predominant ash forming elements are K, Na, Ca, Mg, Fe, Al, Si, P, S, Cl, C, H, and O. The main ash characteristics that can lead to agglomeration come from Na and K, alkali metals. When agglomerates form they disrupt the dynamics in the boiler and if the temperature exceeds the melting temperature of the particles for too long the sticky outer layer of bed material can form permanent bonds (sintering), which can lead to slagging or in worst case scenarios complete defluidization. In a study by (Chirone et al., 2006) they report that bed agglomerates begin to form near burning char because of the higher temperatures and this increases the formation of melt, or the particles stickiness. This is especially problematic because it can lead to the formation of eutectics and the escalation of adhesive forces during sintering. Eutectics formed can have melting points as low as 401 °C and 552 °C for $Na_2S_2O_7$ and $Na_3K_2Fe_2(SO_4)_6$ respectively (Dunnu et al., 2010).



(a) Agglomeration from ash melting.



(b) Agglomeration from ash coating.

Figure 2. Prominent modes of Agglomeration.

Agglomerates are a common occurrence in fluidized beds, but a well controlled boiler can control the fluidazation velocity or temperature well enough to break these clusters apart before sintering takes place. Presented in (Skrifvars et al., 1994; Yan et al., 2003), they discuss how the oxygen concentration in the boiler is linked to the risk of sintering, where regions of higher oxygen content can lead to hot spots. Two mechanisms, flue gas recirculation and air flow, can be used to prevent the likelihood of hot spots. Flue gas recirculation into the boiler bed will help in reducing the oxygen content as well as bed temperature. Maintaining a consistent air flow rate, depending upon the quality of the fuel, for fluidizing the bed material can also help in facilitating a well mixed combustion median in order to reduce hot spots.

A wide range of methods for agglomeration detection, when fired with biomass, have been compared and can be grouped into three categories: on-line detection, experimental methods (controlled agglomeration tests), and theoretical evaluations (fuel ash analysis), (Gatternig, 2015). The last two methods seem reasonable when considering a fuel that is relatively homogeneous, but RDF is a complex and difficult fuel to model and predict melting points, and ash composition because of its composition variability. In most industrial fluidized bed boilers it is routine to take random fuel samples to check for composition consistency. These samples could be used in agglomeration indexes (Visser, 2004; Vamvuka et al., 2008) that look at, among other elements, Na and K in the fuel to determine the likelihood of agglomerate formation. However, this approach assumes that the fuel characteristics will remain the same year after year, but in reality RDF composition is based on the consumers and such an assumption is unrealistic. Also, this method only considers a fraction of the predictability of agglomeration by looking at fuel composition and neglects fuel-ash-bed material interactions, and therefore can give a bit of insight into agglomeration tendencies and should be used cautiously. Another method for agglomeration detection is through advanced multicomponent/multiphase thermodynamic modeling (Lindberg et al., 2013), but there is currently a lack of comprehensive thermodynamic databases for ash compounds and the phases formed during combustion.

Therefore, online detection is the most suitable way to predict agglomeration when RDF is used. Current methods used for the early detection of agglomeration have been explored by looking at pressure drops and temperature fluctuations. However, at this point, the probability of defluidization is already prevalent and leaves no options for the operator but to shutdown the plant, and is not a suitable form of early detection. An alternative method, suitable for early detection, of agglomerates on a small-scale in fluidized beds has been proposed by (Nijenhuis et al., 2007). They were able to develop an early agglomeration recognition system that detects very small changes in hydrodynamic multiphase systems, and allow detection of agglomeration up to 60 minutes before occurring.

The method proposed by the authors in this paper for the detection of agglomerates is based on the minimum fluidization needed to fluidize the bed material. As the diameter of the bed material increases so will the amount of air need to fluidized the bed material in order to operate the boiler within the proper temperature limits. Hence modeling the minimum fluidization can be used in early warning detection of minimal to extreme agglomerate formation.

3 Methodology

The dynamic model used has been calibrated and validated using data from a RDF fired CFB, boiler 6, at MälarEnergi, in Västerås, Sweden. It was possible to back calculate the mass flow rate of fuel by conducting a heat balance on the boiler's heat exchangers. By this method it was determined that 16.8 kg/s of RDF are fed into the boiler, taking the thermal efficiency of the boiler into consideration, this corresponds to MälarEnergi's reported value of 30 tonne/hour. The model is designed to allow for the real input of primary air, secondary air, and flue gas recirculate mass flows and corresponding temperatures. It is assumed that the mass flow rate of the fuel and its respective LHV are constant.

3.1 Description of Model

A model has been designed in DYMOLA using Modelica programming language. The reason for having a dynamic model is to be able to capture the transient behavior of RDF through the combustion process. With the end goal of being able to have a model that has the ability to not only monitor agglomeration, but to also be used for emissions tracking, decision support, and fault detection. The Modelica modeling language allows users to build model libraries with ability to reuse component blocks and to easily change parameters to match any complex dynamic system.

Individual component blocks were made for the CFB-loop, also represented in Figure 1, for the boiler bed, freeboard, cyclone, and hot cycle recirculate (super heater (SH)) respectively. Multiple functions have been been written in order to accurately represent the thermodynamic properties for all in-coming streams, bottom and fly ash, bed material,and flue gas.

3.2 Mass and Energy Balances

The model is based on mass and energy balances (equations 1 and 2) used for the freeboard and bed of the boiler, cyclone, and superheater following that of a similar approach to that presented in (Basu and Fraser, 2015; Gungor, 2009). Where i represents the control volume in the CFB-loop (A, B, C, D) in Figure 1, m is the mass, H is the enthalpy, α is the percentage of combustion, and $\dot{Q}$ is the heat released during combustion.

$$\frac{d(m_i)}{dt} = \Sigma \dot{m}_{in,i} - \Sigma \dot{m}_{out,i} \tag{1}$$

$$\frac{d(m_i H_i)}{dt} = \dot{m}_{in,i} H_{in,i} - \dot{m}_{out,i} H_{in,i} + \alpha \dot{Q}_{released,i} \quad (2)$$

It was possible to calculate the enthalpies of all constituent parts in the energy balance, equation 3, where T_i is the temperature in the control volume and T_{ref} is reference temperature taken at 289.15K.

$$\frac{d(H_i)}{dt} = Cp_i(T_i - T_{ref}) \quad (3)$$

The fuel composition and flow rate is constantly changing, and in turn so is the boiler bed and flue gas temperature, so a series of functions have been developed to calculate the specific heat for each element in the control volume in the energy balance. In this way, instead of assuming constant values, the model has the capability of more accurately simulating temperature profiles for the bed and freeboard of the boiler. The function developed for calculating the specific heat (Cp) of the gas follows the practices of (Wester, 1987), equation 4.

$$Cp = \frac{1}{T - T_{ref}} \int_{T_{ref}}^{T} \sum_{n=-1}^{n=7} a_c^n \left(\frac{T}{1000} \right)^n \quad (4)$$

The coefficients for a_c^n represent air, N_2, O_2, CO_2, and water vapor. The gas thermophysical properties and composition were calculated by using flue gas stoichiometric calculations. The flue gas composition is found to primarily be comprised of CO_2, H_2O,O_2, and N_2. Where the model uses the ultimate analysis of the fuel as input, calculates the amount of excess air in the system, and then the flue gas composition can be used to calculate the specific heat, equation 5, as well as viscosity and density at the operating temperature.

$$\begin{aligned} Cp_{gas} = {} & [N_2]Cp_{N_2} + [O_2]Cp_{O_2} \\ & + [CO_2]Cp_{CO_2} + [H_2O]Cp_{H_2O} \end{aligned} \quad (5)$$

The solids contained within each control volume consist of fuel, sand, ash, and char, as shown in equation 6. Where α, β, γ, and δ represent the percentage of fuel, sand, ash, and char respectively that are in the control volume i.

$$C_{p,solids,i} = \alpha \cdot C_{p,fuel} + \beta \cdot C_{p,sand} + \gamma \cdot C_{p,ash} + \delta \cdot C_{p,char} \quad (6)$$

The functions for fly and bottom ash were developed by using chemical compositions found in (Chang et al., 1997). Where they analyzed the chemical composition and determined the compounds and corresponding percent composition of ash, and found that CaO, SiO_2, Al_2O_3, Fe_2O_3, ZnO, MgO, Cr_2O_3 were the major components (in decreasing order). Specific heat correlations for the compounds were developed from (Abu-Eishah et al., 2004; Karditsas and Baptiste, 1995; Madelung et al., 1999), therefore temperature dependent ash specific heat values could be estimated, equation 7.

$$C_{p,ash,i} = \Sigma \left(\xi_j \cdot C_{p,j} \right) \quad (7)$$

where i represents either fly or bottom ash, ξ is the compound percentage in the ash, and j is the compound. With the calculation of Cp_{gases} and Cp_{solids} and corresponding mass flows equations 2 and 3 can be used to determine the boiler temperature.

3.3 Hydrodynamics

CFBs operate at a higher gas velocity and therefore the gas and solids within the boiler act like a fluid. In the bed of a CFB boiler there is a dense emulsion phase, where the gas moves through the inventory as large bubbles. At the top of the bed, where the secondary air input is, the bubbles burst dispersing inventory and unburnt char into the freeboard. The addition of this secondary air, theoretically, allows for a dilute well-mixed phase throughout the freeboard. The freeboard is typically further segregated into a core and annulus. The core is where the gas and solids flow upward, fine particles are carried out of the bed, and coarser particles tend to form clusters and then fall back down in the annulus region as a thin film, where the further up in the freeboard the smaller the annulus region becomes, as shown in Figure 1.

The minimum amount of air velocity required to fluidized the bed material is called the minimum fluidization velocity u_{mf} and is the velocity at which the drag of the fluidized median (sand) is equal to the weight of the bed material. Following the method used in (Kunii and Levenspiel, 2013) the minimum fluidization can be determined, equation 8.

$$\begin{aligned} & \frac{1.75}{\varepsilon_{mf}^3 \phi} \cdot \left(\frac{d_p U_{mf} \rho_g}{\mu_g} \right)^2 + \frac{150(1-\varepsilon_{mf})}{\varepsilon_{mf}^3 \phi^2} \cdot \frac{d_p U_{mf} \rho_g}{\mu g} = \\ & \frac{d_p^3 \rho_g (\rho_p - \rho_g) g}{\mu_g^2} \end{aligned} \quad (8)$$

Where ε_{mf} is the void fraction at minimum fluidizing conditions, ϕ is the sphericity of the sand, d_p is the average diameter of sand particles, ρ is either the density of the fluidizing median or sand, μ_g is the viscosity of the fluidizing median, and g is gravity.

As the amount of air is increased into the boiler the minimum fluidazation velocity of the gas will reach the superficial velocity (U_s) and is affected by the amount of gas, density and size in a given control area. Typically values for industrial CFBs is in the range of 5-10 m/s, (Basu and Fraser, 2015). This velocity carries bed material, ash, and unburnt char up into the freeboard, where a portion of these circulate within the boiler but there is also a net solid flux, G_s that leaves the boiler to be recirculated back through the loop-seal. Where G_s is equal to the amount

of solids going up in the boiler minus the solids circulating within the boiler. A correlation between G_s and operational conditions was presented by (Guan et al., 2010), equation 9, where G_s is in the rang of 200 and 400 kg/m^2, which is within the operating limits of the designed CFB.

$$\frac{G_s d_p}{\mu_g} = 547 Ar^{0.248} \left(\frac{U_s}{\sqrt{gD}} \right)^{0.375} \left(\frac{D}{H} \right)^{0.195} \quad (9)$$

Where Ar is Archimedes number, D is the diameter of the freeboard, and H is the height of the freeboard.

4 Results and Discussion

4.1 Validation

The model has been constructed to predict the performance of a 160 MW CFB boiler. The model has been designed to predict the bed and flue gas temperatures. The validity of the model has been determined by comparing data from boiler 6 at MälarEnergi. Figure 3 illustrates a simulation, for a week, of the flue gas temperature as it exits to the cyclone. The accuracy of the model is quite good initially, less than a few percent. However, it can be seen that when there is sudden increase or decrease the model tends to underestimate, or overestimate, the respective temperature. Since the model is currently designed using a constant fuel mass flow rate and heating value it is reasonable to believe that this is attributing to the deviation in the predicted temperature. The model's profile is able to follow that of the actual profile, but if the quality of the fuel coming into the boiler is poor this would be reflected in a substantial drop in the temperature profile.

4.2 Agglomerate Prediction

There is a lot in the literature about methods used to predict CFB failure from agglomeration, but typically these do not take into account the impact of the combustion environment (Yan et al., 2003) like gas to particle interactions and alkali vapor condensing.

Agglomeration prediction has been studied to a lengthy extent where fossil and biomass fuels are concerned. The standard procedure in industry for predicting agglomeration is determined from pressure drops and temperature changes (Gatternig, 2015), but it has already been mentioned that this is not suitable for early detection. For early detection, using plant process parameters, it is possible to detect agglomeration by looking at the minimum fluidization velocity, where a small increase shows that the diameter of the circulating bed material is increasing.

This is because with the onset of agglomerates and then sintering, the minimum fluidization required to fluidized the bed material will slowly increase. Therefore, agglomeration can be monitored by modeling process parameters, while operators can still keep an eye on pressure drops in the boiler bed and temperature fluctuations.

As mentioned the composition of RDF can vary from hour to hour. However, the composition of sand, with an average diameter in the range of $0.40 < d_p < 0.63$, is known and as agglomerates form this will require a higher minimum fluidization. Figure 4 confirms the presence of a relationship between the bed particle size increasing and the required increase in the minimum fluidization. Where the bed material size distribution was obtained from Mälarenergi, daily average, and the minimum fluidization is the model's prediction, taken on a daily average. This method allows for the detection of a a small change in the bed material average diameter with a suitable amount of time for the operator to make the decision to add fresh sand to the boiler before there is a subsequent formation of further agglomeration, slagging, or possibly complete defluidization.

Figure 5 illustrates the main elemental composition in g/kg TS of the bed material samples over the week in question. Because the ash melts are coating the sand particles it is reasonable that the main composition comes from silica, with an average value of 301 g/kg TS. However, a portion of the silica could come from glass fines in the fuel mix, but this value is unknown. It should also be noted that all of these elements, as stated before, are predominant ash forming elements. Therefore, it can be quantified that the agglomeration of the sand has occurred due to either ash melting or ash coating.

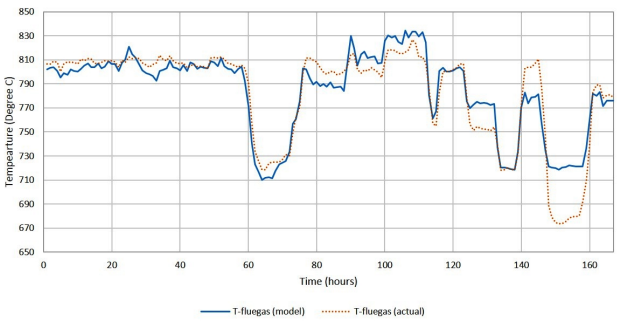


Figure 3. Temperature profile of the fluegas during the study period.

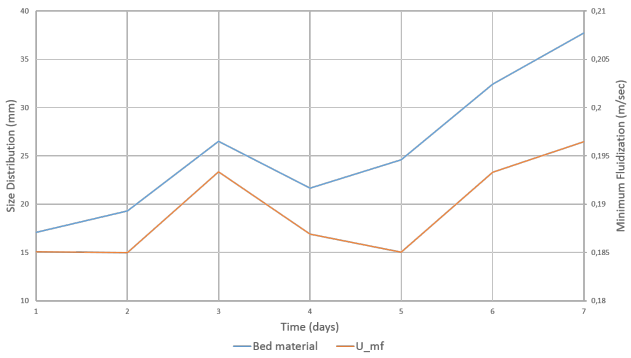


Figure 4. Bed Material Size Distribution $0.4 < d_p < 0.63$ mm Correlates to the Minimum Fluidization.

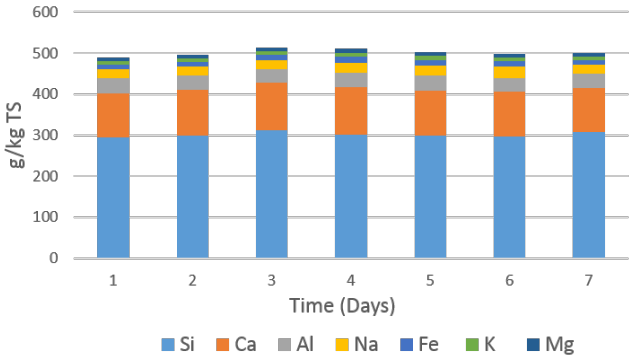


Figure 5. Primary elemental composition of the bed material during the study period.

5 Conclusions

The model presented shows the ability to determine the agglomeration of bed material while off-line. This has been accomplished by building a dynamic model using process parameters as input to model and calculate the minimum fluidization velocity required to maintain a boiler operating temperature within the range of 750 °C and 900 °C. It should be noted that during the seven days there were periods where the fluegas temperature dropped below 750 °C and can be attributed to the possibility of poor fuel. The simulated results show that an increase in minimum fluidization velocity corresponds to an increase in bed material share that is greater than the average diameter of the sand used, $0.40 < d_p < 0.63$. Since agglomeration is prevalent no mater what the fuel source is, it is possible to implement this model as a means for real-time detection of agglomerate formation as a means for decision support to operators.

Compared to other CFB models. The library developed in this study can potentially be reused for any CFB installation through the ease of the drag-and-drop nature of object-oriented programming. The model presented has the ability to handle the transient behavior of RDF, not only for modeling agglomeration, but to also capture the dynamics of the temperature fluctuation in the boiler, hence the possibility for real-time emissions monitoring. Decision support and fault detection can also be implemented by formulating a probabilistic distribution through Bayesian nets. Only one model is needed to monitor multiple aspects of the combustion process in CFBs.

Acknowledgment

The author would like to thank Linda Svensson, Allmyr Marianne, and Lisa Granström for their assistance in data acquisition and process information on block 6 at Mälarenergi, Västerås. The authors would also like to thank Erik Dahlquist at Mälardalen University for his leadership and guidance. Funding for this work comes from The PolyPo Project within the Future Energy Research Profile at Mälardalen University.

Nomenclature

Acronyms

CFB	circulating fluidized bed
RDF	Refuse derived fuel

Greek Symbols

ε_{mf}	Void fraction at minimum fluidization
μ	Viscosity, $\frac{kg}{ms}$
ϕ	Sphericity
ρ	density, $\frac{kg}{m^3}$

Roman Symbols

Ar	Archimedes number, $\frac{\rho_p(\rho_p-\rho_g)gd_p^3}{\mu^2}$
Cp	Specific heat, $\frac{kJ}{kgK}$
d_p	Particle diameter, m
g	Gravity, $\frac{m}{s^2}$
G_s	Net solid flux, $\frac{kg}{m_s^2}$
H	Enthalpy, $\frac{kJ}{kg}$
h	Height of the freeboard, m
m	mass, kg
Q	Heat released, KW
T	Temperature, K
U_{mf}	Minimum fluidization velocity, $\frac{m}{s}$
U_s	Superficial Velocity, $\frac{m}{s}$

Subscripts

g	Gas
i	Corresponding control volume
p	Particle
ref	Reference temperature, ambient

References

S. I. Abu-Eishah, Y. Haddad, A. Soliman, and A. Bajbouj. A new correlation for the specific heat of metals, metal oxides and metal fluorides as a function of temperature. *Latin American Applied Research*, 34(OCTOBER 2004):257–265, 2004. ISSN 03270793.

Cyril Jose E. Bajamundi, Pasi Vainikka, Merja Hedman, Jaani Silvennoinen, Teemu Heinanen, Raili Taipale, and Jukka Konttinen. Searching for a robust strategy for minimizing alkali chlorides in fluidized bed boilers during burning of high SRF-energy-share fuel. *Fuel*, 155(2015):25–36, 2015. ISSN 00162361. doi:10.1016/j.fuel.2015.03.087.

Prabir Basu and S A Fraser. *Circulating Fluidized Bed Boilers: Design, Operation and Maintenance*. Springer International Publishing, Switzerland, doi: 10.10 edition, 2015. ISBN 9783319061726. doi:10.1007/978-3319061733.

Ni-Bin Chang, H.P. Wang, and K.S. Lin. *Comparison between MSW ash and RDF ash from incineration process*. Dec 1997.

Riccardo Chirone, Francesco Miccio, and Fabrizio Scala. Mechanism and prediction of bed agglomeration during fluidized bed combustion of a biomass fuel: Effect of the reactor scale. *Chemical Engineering Journal*, 123(3):71–80, 2006. ISSN 13858947. doi:10.1016/j.cej.2006.07.004.

Gregory Dunnu, Jörg Maier, and Günter Scheffknecht. Ash fusibility and compositional data of solid recovered fuels. *Fuel*, 89(7):1534–1540, 2010. doi:10.1016/j.fuel.2009.09.008.

A. L. Elled, L. E. Åmand, and B. M. Steenari. Composition of agglomerates in fluidized bed reactors for thermochemical conversion of biomass and waste fuels: Experimental data in comparison with predictions by a thermodynamic equilibrium model. *Fuel*, 111:696–708, 2013. ISSN 00162361. doi:10.1016/j.fuel.2013.03.018.

Bernhard Gatternig. *Predicting Agglomeration in Biomass Fired Fluidized Beds*. PhD thesis, 2015.

Guoqing Guan, Chihiro Fushimi, and Atsushi Tsutsumi. Prediction of flow behavior of the riser in a novel high solids flux circulating fluidized bed for steam gasification of coal or biomass. *Chemical Engineering Journal*, 164(1):221–229, 2010. ISSN 13858947. doi:10.1016/j.cej.2010.08.005.

Afsin Gungor. One dimensional numerical simulation of small scale CFB combustors. *Energy Conversion and Management*, 50(3):711–722, 2009. ISSN 01968904. doi:10.1016/j.enconman.2008.10.003.

Francisco D. Hernandez-Atonal, Changkook Ryu, Vida N. Sharifi, and Jim Swithenbank. Combustion of refuse-derived fuel in a fluidised bed. *Chemical Engineering Science*, 62(1-2):627–635, 2007. ISSN 00092509. doi:10.1016/j.ces.2006.09.025.

Panayiotis Karditsas and Marc-Jean Baptiste. Thermal and structural properties of fusion related materials, 1995.

Daizo Kunii and Octave Levenspiel. *Fluidization engineering*. Elsevier, 2013.

Weigang Lin, Kim Dam-Johansen, and Flemming Frandsen. Agglomeration in bio-fuel fired fluidized bed combustors. *Chemical Engineering Journal*, 96(1-3):171–185, 2003. ISSN 13858947. doi:10.1016/j.cej.2003.08.008.

Daniel Lindberg, Rainer Backman, Patrice Chartrand, and Mikko Hupa. Towards a comprehensive thermodynamic database for ash-forming elements in biomass and waste combustion: Current situation and future developments. *Fuel Processing Technology*, 105:129–141, 2013. doi:10.1016/j.fuproc.2011.08.008.

Zhen-Shu Liu, Tzu-Huan Peng, and Chiou-Liang Lin. Effects of bed material size distribution, operating conditions and agglomeration phenomenon on heavy metal emission in fluidized bed combustion process. *Waste management (New York, N.Y.)*, 32(3):417–25, 2012. ISSN 1879-2456. doi:10.1016/j.wasman.2011.10.033.

O. Madelung, U. Rössler, and M. Schulz. In *II-VI and I-VII Compounds; Semimagnetic Compounds*, volume 41B of *Landolt-BÃűrnstein - Group III Condensed Matter*. 1999. ISBN 978-3-540-64964-9.

J. Nijenhuis, R. Korbee, J. Lensselink, J. H a Kiel, and J. R. van Ommen. A method for agglomeration detection and control in full-scale biomass fired fluidized beds. *Chemical Engineering Science*, 62(1-2):644–654, 2007. ISSN 00092509. doi:10.1016/j.ces.2006.09.041.

Anita Pettersson, Fredrik Niklasson, and Farzad Moradian. Reduced bed temperature in a commercial waste to energy boiler - Impact on ash and deposit formation. *Fuel Processing Technology*, 105:28–36, 2013. ISSN 03783820. doi:10.1016/j.fuproc.2011.09.001.

Georgy Ryabov, Dmitry Litoun, and Eduard Dik. Agglomeration of bed material: Influence on efficiency of biofuel fluidized bed boiler. *Thermal Science*, 7(1):5–16, 2003. ISSN 0354-9836. doi:10.2298/TSCI0301005R.

Fabrizio Scala and Riccardo Chirone. Fluidized bed combustion of alternative solid fuels. *Experimental Thermal and Fluid Science*, 28:691–699, 2004. ISSN 08941777. doi:10.1016/j.expthermflusci.2003.12.005.

Fabrizio Scala and Riccardo Chirone. An SEM/EDX study of bed agglomerates formed during fluidized bed combustion of three biomass fuels. *Biomass and Bioenergy*, 32(3):252–266, 2008. ISSN 09619534. doi:10.1016/j.biombioe.2007.09.009.

Bengt-Johan Skrifvars, Mikko Hupa, Rainer Backman, and Matti Hiltunen. Sintering mechanisms of FBC ashes. *Fuel*, 73(2):171–176, 1994.

Wolfgang Staber, Sabine Flamme, and Johann Feltner. Methods for determining the biomass content of waste. *Waste management & research : the journal of the International Solid Wastes and Public Cleansing Association, ISWA*, 26(1):78–87, 2008. ISSN 0734-242X. doi:10.1177/0734242X07087313.

D Vamvuka, D Zografos, and G Alevizos. Control methods for mitigating biomass ash-related problems in fluidized beds. *Bioresource technology*, 99(9):3534–44, 2008. ISSN 0960-8524. doi:10.1016/j.biortech.2007.07.049.

H J M Visser. The influence of Fuel Composition on Agglomeration Behaviour in Fluidised-Bed Combustion. *ECN Biomass*, (September):44, 2004.

Joachim Werther. Fluidized-Bed Reactors. In *Ullmann's Encyclopedia of Industrial Chemistry*. 2007. ISBN 3527306730. doi:10.1002/14356007.b04_239.pub2.

L. Wester. *Tabeller och diagram för energitekniska beräkningar*. L. Wester, 1987.

Rong Yan, David Tee Liang, Karin Laursen, Ying Li, Leslie Tsen, and Joo Hwa Tay. Formation of bed agglomeration in a fluidized multi-waste incinerator. *Fuel*, 82(7):843–851, 2003. doi:10.1016/S0016-2361(02)00351-4.

dSPACE Implementation for Real-Time Stability Analysis of Three-Phase Grid-Connected Systems Applying MLBS Injection

Tomi Roinila[1] Roni Luhtala[1] Tommi Reinikka[1] Tuomas Messo[2] Aapo Aapro[2] Jussi Sihvo[2]

[1]Department of Automation Science and Engineering, Tampere University of Technology, Finland
{tomi.roinila}@tut.fi
[2]Department of Electrical Engineering, Tampere University of Technology, Finland {tuomas.messo}@tut.fi

Abstract

Renewable resources such as solar and wind are most commonly connected to a utility grid through inverters. The stability and system characteristics of such systems can be defined by the ratio of grid impedance to the inverter output impedance. Since the impedances vary over time with numerous operation conditions, real-time measurements are required to verify the stability. The impedance measurement technique based on maximum-length-binary-sequence (MLBS) injection and Fourier techniques has been proven to be an efficient option for online analysis of grid-connected systems. This paper shows how a hardware-in-the-loop simulation based on dSPACE can be implemented for stability analysis of a grid-connected inverter using the MLBS injection. The method makes it possible to modify the inverter control characteristics and grid conditions online, thereby providing means for various stability and control design studies for grid-connected systems. We have presented a measurement example based on a three-phase grid-connected inverter and used this example to demonstrate the implementation.

Keywords: frequency response, power system measurements, spectral analysis, signal design, real-time systems

1 Introduction

The most common way to connect renewable resources such as wind turbines or photovoltaic generators to a power grid is through inverters. As the penetration of the inverter-connected systems increases, it has globally important effect on the grid's performance. Consequently, interaction issues between the grid-parallel inverters and power grids have been topics of extensive research in recent years (Cespedes and Sun, 2014b; Lu et al., 2015; Hu et al., 2015). One of the most important topics has been the harmonic resonance generated due to the mismatch between the inverter's output impedance and grid impedance. This is commonly known as the harmonics-related power-quality problem, which has had a significant effect on overall energy efficiency and even grid stability (Sun, 2011). Recent studies have shown that the instability can be avoided by measuring the impedances of the grid and inverter, and based on the measurements, adaptively changing the inverter parameters (Cespedes and Sun, 2014a).

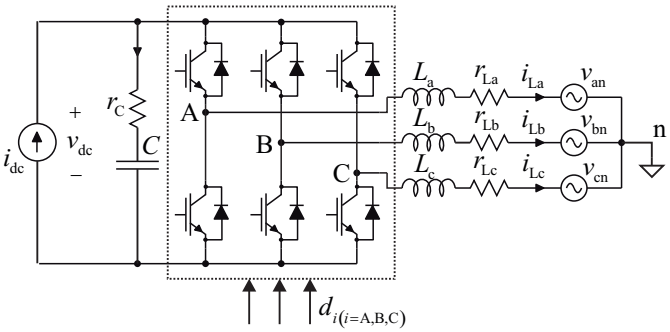


Figure 1. Grid-connected renewable energy inverter.

The impedance measurements of grid-connected systems using a broadband excitation and Fourier techniques have become a popular method in recent years (Barkley and Santi, 2009; Cespedes and Sun, 2014a; Roinila et al., 2013). This method involves injecting an external current on top of the normal output current of the inverter or grid, measuring the resulting voltage responses, and applying Fourier analysis to extract the frequency components in both the voltage and current. The grid or inverter impedance is then determined by the ratio between the voltage and current at different frequencies. The most common excitation types have been impulse (Cespedes and Sun, 2014a) and maximum-length binary sequence (MLBS) (Roinila et al., 2014), from which the MLBS has shown superiority over the impulse. The MLBS is a deterministic and periodic signal. Hence, the effect of external noise can be computationally reduced, and multiple periods can be applied through spectral averaging to further increase the signal-to-noise ratio (SNR). As a result, the amplitude of the excitation can be kept at a much lower level than the amplitude of many other types of excitations. Due to the binary form of the MLBS, the injection is extremely easy to implement, even with a low-cost application, the output of which can only cope with a small number of signal levels.

This paper considers real-time impedance measurements of a grid-connected system using hardware-in-the-loop (HIL) simulation based on dSPACE software

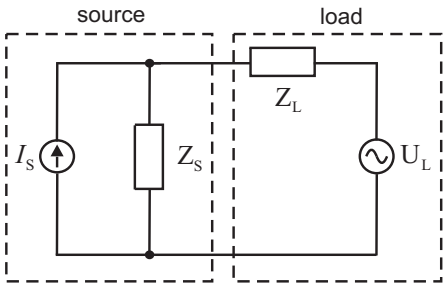


Figure 2. Interconnected source-load subsystem.

and Matlab/Simulink. dSPACE is widely used in real-time analysis and control of various power-electronics applications including three-phase grid-connected systems (Ghani et al., 2009), digitally controlled power converters (Monti et al., 2003), and back-to-back converters (Deshpande et al., 2012). Matlab/Simulink provides a functionality that generates a C-code from the Simulink model. Using dSPACE's real-time interface (RTI), the C-code can be automatically implemented into the I/O board of dSPACE. This method makes it possible to modify the grid characteristics and the inverter's controller parameters online, which in turn enables various stability and control design studies for grid-connected systems in time-varying conditions.

This paper will show the implementation of a real-time impedance measurement of a grid-connected system using dSPACE. No external signal generator or data-acquisition units are required; the signal generation, injection and computations are all performed in dSPACE. We show the implementation steps, starting from generating the MLBS. As the paper does not consider dSPACE in detail, the reader should have a basic knowledge of the software.

The remainder of this paper is organized as follows. Section 2 briefly reviews the theory behind the stability analysis of grid-connected systems and the synthesis of the MLBS. Section 3 gives an example of a grid-connected system operated by dSPACE, and provides guidelines for generating the MLBS and obtaining the system characterizing responses. Section 4 draws conclusions.

2 Theory

2.1 Stability of Grid-Connected System

Figure 1 depicts a three-phase inverter for direct interfacing of a photovoltaic generator. The inverter is comprised of six power electronic switches, a DC capacitor, and threephase inductors. The inverter controls its switches between conducting and non-conducting mode with sinusoidal control voltages in order to transform DC from the renewable energy source into the three-phase AC required by the power grid.

The stability of a inverter-connected system can be easily assessed in the frequency domain by constructing a small-signal state-space representation for the interfacing inverter and the load subsystem. The stability analysis can

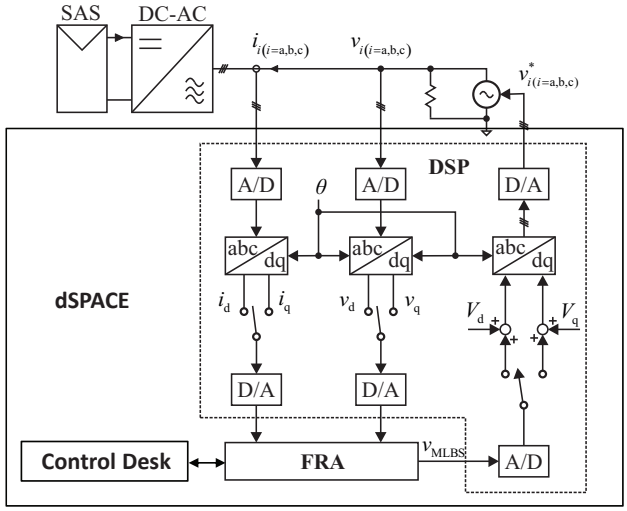


Figure 3. Circuit diagram of three-phase grid-connected inverter connected to real-time spectrum analyzer based on MLBS.

be conducted by applying the Nyquist stability criterion to the impedance ratio in an interface (Wang et al., 2014).

Figure 2 shows a simple example of a single-phase system in which the system consists of one source powering a single load. The source is modeled by a Norton equivalent circuit, as a current source I_S in parallel with the source impedance Z_S. The load voltage is denoted by U_L and the load impedance by Z_L. This combination applies for a grid-parallel inverter in which the grid acts as a voltage-type load and the inverter resembles a controlled current source. Assuming that the source is stable when unloaded and that the load is stable when powered by an ideal source, the stability and other dynamic characteristics of the interconnected system can be determined from the transfer function

$$G(s) = \frac{1}{1 + Z_L(s)/Z_S(s)} \tag{1}$$

The interconnected system is only stable if the impedance ratio $Z_L(s)/Z_S(s)$ satisfies the Nyquist stability criterion. Power systems that are more complex can be represented in the same form and analyzed similarly by putting together multiple sources into a source subsystem and loads into a load subsystem. In general, a grid-connected inverter does not suffer from resonance phenomena caused by impedance-based interactions if the output impedance of the inverter is shaped so that it has a larger magnitude than grid impedance at every frequency.

Three-phase inverters can be modeled in the DQ domain by using direct (d) and quadrature (q) components (Yazdani and Iravani, 2010). The output impedance can be represented in the matrix form shown in (2). The cross-coupling impedances Z_{qd} and Z_{dq} can usually be neglected in stability analysis because they are typically very small in magnitude. Analogous to single-phase systems, the stability of a three-phase system can be determined from the transfer functions in (3) and (4) by applying the Nyquist

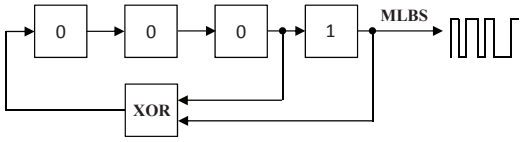


Figure 4. Shift register with XOR and feedback.

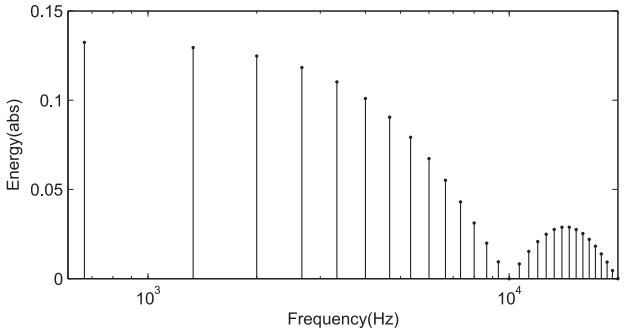


Figure 5. Power spectrum of 15-bit-length MLBS generated at 10 kHz.

stability criterion. Both impedance ratios have to satisfy this criterion for stable operation.

$$\mathbf{Z}_{\mathrm{S}} = \begin{bmatrix} Z_{\mathrm{d}} & Z_{\mathrm{qd}} \\ Z_{\mathrm{dq}} & Z_{\mathrm{q}} \end{bmatrix} \tag{2}$$

$$G_{\mathrm{d}}(s) = \frac{1}{1 + Z_{\mathrm{L}}(s)/Z_{\mathrm{d}}(s)} \tag{3}$$

$$G_{\mathrm{q}}(s) = \frac{1}{1 + Z_{\mathrm{L}}(s)/Z_{\mathrm{q}}(s)} \tag{4}$$

2.2 Maximum-Length Binary Sequence

Pseudo-random binary sequence (PRBS) is a periodic broadband signal based on a sequence of length N. The most commonly used signals are based on maximumlength sequences (MLBS). Such sequences exist for $N = 2^n - 1$, where n is an integer. These are popular because they can be generated using feedback shift-register circuits. (Godfrey, 1993)

Figure 4 shows an example of a shift-register circuit for generating an MLBS of a length $2^4 - 1 = 15$. The feedback is generated from stages 3 and 4. The register can be started with any number other than 0,0,0,0. In practice, the values 0 and 1 are mapped to -1 and +1 to produce a symmetrical MLBS with an average close to zero.

Figure 5 shows the form of the power spectrum of the MLBS generated by the shift register shown in Figure 4. The sequence is generated at 10 kHz and has signal levels ± 1 V. The power spectrum has an envelope and drops to zero at the generation frequency and its harmonics. The MLBS x has the lowest possible peak factor $|x|_{\mathrm{peak}}/x_{\mathrm{rms}} = 1$ regardless of its length, which means that the sequence is well suited to sensitive systems that require small-amplitude perturbation. .

Due to the deterministic nature of the sequence, the signal can be repeated and injected precisely and the SNR can be increased by synchronous averaging of the response periods.

3 Implementation in dSPACE

This section provides the main steps and guidelines for the frequency-response-measurement procedure using dSPACE. The steps are shown through an example in which the output impedance is measured from a threephase grid-connected inverter.

3.1 System Setup

Figure 3 shows the setup of the system under study. The goal is to measure the d- and q-components of inverter's output impedance. A similar approach can be used to measure the grid impedance or any other system-characterizing frequency response, but the example only considers the output-impedance measurement. The system comprises the power stage and real-time frequency-response analyzer (FRA) based on the MLBS injection. The powerstage components are the photovoltaic generator, the inverter, and the three-phase grid emulator. The details of the power-stage components are omitted because they are not within the scope of this paper.

The MLBS is implemented in dSPACE and runs parallel with the inverter control functions. The "Control Desk"-block in Figure 3 depicts a PC, which is used to modify the MLBS parameters such as the length of the injected signal and its amplitude. The MLBS is injected to the d- or q-component of the grid reference voltage. The perturbed three-phase currents and voltages of the inverter are transformed into their corresponding d- and q-components and collected by the FRA.

The measurements of the d- and q-components of the impedance require two separate measurement cycles; one for injecting and collecting d-components and one for q-components. It should be emphasized that various transfer functions can be measured without disconnecting the system. The possible variables can be defined in dSPACE and can be switched in the "Control Desk" -block. For example, the effect of different control loops on a specific transfer function can easily be analyzed online.

Figure 6 is a diagram of the real-time spectrum analyzer including the generation of the MLBS, the sequence injection, and data collection. All the rectangular boxes denote the Simulink blocks that are used for dSPACE. The shift register is implemented by unit-delay blocks. The output of exclusive-or (XOR) replaces the first bit of the sequence through feedback. The generation frequency of the injection is set by the delay value of the unit-delay blocks.

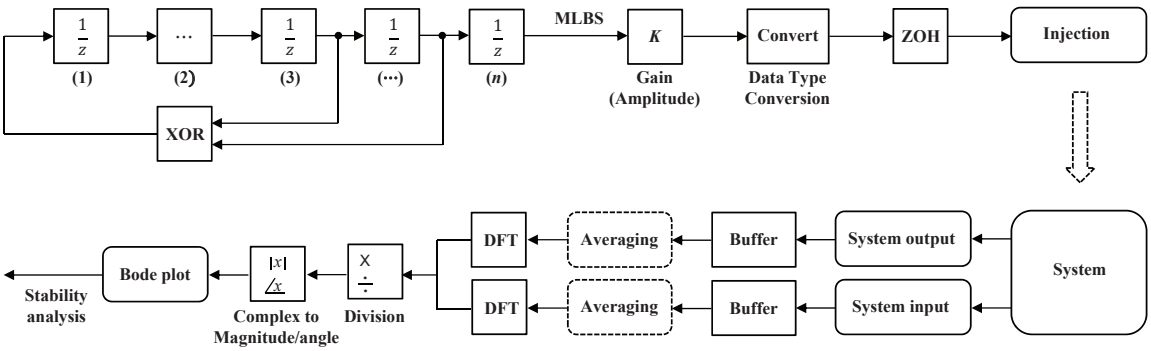


Figure 6. Diagram of the excitation generation, injection and data collection.

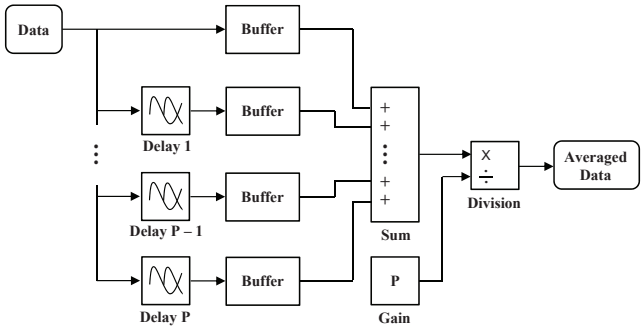


Figure 7. Diagram of the averaging procedure of measured data.

The MLBS is amplified by an adjustable gain (K), after which the sequence is converted from logical numbers to realworld numbers. The zeroorder-hold block is required by dSPACE. The presented concept allows continuous and repeating generation and injection of the MLBS into the system.

The presented implementation also makes it possible to change the injection amplitude in real time. Hence, depending on the noise level and nonlinearities, the amplitude can be experimentally adjusted so that the produced injection energy is high enough. The injection amplitude cannot be too high because the grid-connected systems are typically highly sensitive to external signals and nonlin-earities may easily arise.

The measurements of input and output data are continuously collected and buffered. Once the data is buffered a DFT matrix is applied to perform the Fourier transform (Sundararajan, 2001). The reason for the use of the DFT matrix is that the length of the buffered data is $2^n - 1$ (length of the MLBS). A readily available FFT-block could be used but the block only accepts a data sequence of length 2^n. Therefore, the fast Fourier transform is not applied in the implementation.

The Fourier transformed output data is divided by the input data resulting in the complex transfer function. The Bode plot is obtained by computing the magnitude and phase from the complex data. The refresh rate of the Bode plot is $2^n/f_\mathrm{s}$, where n is the length of the shift register and f_s is the sampling frequency.

The effect of external noise can be reduced by applying averaging. Figure 7 shows the diagram for moving average. Input and output data are delayed by $i \cdot f_\mathrm{s} 2^n$, where $i = 1, 2, \ldots, P$ where P denotes the number of injection periods.

3.2 Experiment

The applied MLBS injection is generated through a 7-bit-length shift register, resulting in an 127-bit-long sequence. The sampling frequency f_s and injection generation frequency f_g are set to 8 kHz and 4 kHz, respectively. Using the specified values for injection length and generation frequency, the frequency resolution is fixed to $f_\mathrm{g}/2^n = 4\mathrm{kHz}/127 \approx 31\mathrm{Hz}$. The measurement system is built in Matlab/Simulink as shown in Figure 6, after which the model is transferred to dSPACE as C-code.

Figs. 8 and 9 show a sample measurement of the d- and q-components of the inverter's output impedance. The curves are averaged over 12 injection periods. Hence, the time for one measurement cycle took approximately 0.38 s. Because a moving average was applied (Figure 7), a new Bode plot was obtained after each injection period. Therefore, the refresh rate of the Bode plot was approximately 31 Hz. The reference responses are obtained by sine sweeps. Due to long measurement time of the sine-sweep technique (approximately 10 min in the application), the method cannot be applied in practice. The figure shows that the results obtained by the MLBS accurately follow the reference, showing only a few decibels and degrees of error. No external data-acquisition devices were used in the experiment. The signal generation, injection and computations were all performed in dSPACE.

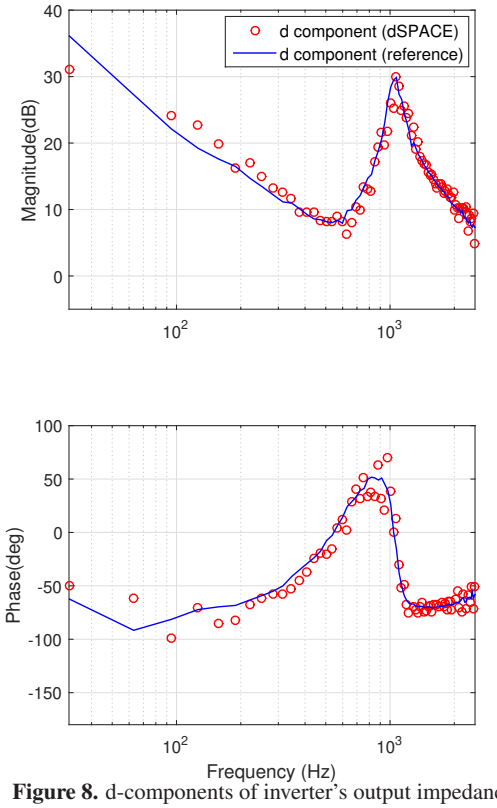


Figure 8. d-components of inverter's output impedance.

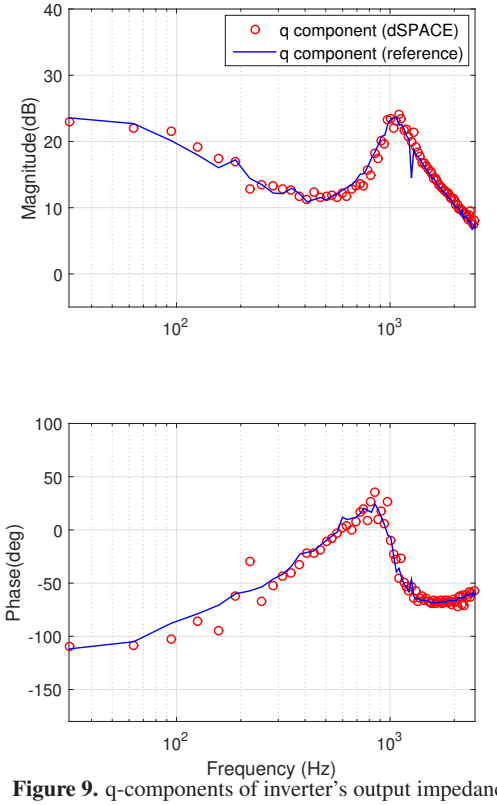


Figure 9. q-components of inverter's output impedance.

4 Conclusions

This paper has presented methods for real-time stability analysis of grid-connected systems using hardware-in-the-loop simulation based on dSPACE and Matlab/Simulink. The MLBS broadband injection was used to measure the output impedance of a three-phase grid-connected inverter. Due to the low peak factor, the amplitude of the injection can be kept relatively small compared to other types of excitations, thereby guaranteeing normal system operation during the identification. Other advantages include straightforward generation of the sequence.

A dSPACE implementation for generating the injection and data acquisition was shown. The proposed methods allow continuous monitoring of system performance and real-time adjustments of the injection properties. The method also makes it possible to modify the inverter control characteristics and grid conditions online, which provides means for analysis under time-varying conditions. Due to fast injection and measurement time, the presentedmethod is useful for various on-line and real-time measurements in, for example, adaptive control of three-phase inverters, and stability analysis of grid-connected systems.

Acknowledgements

This work is supported by the Academy of Finland.

References

A. Barkley and E. Santi. Online monitoring of network impedances using digital network analyzer techniques. In *Proc. Applied Power Electronics Conference and Exposition*, pages 440–446, 2009.

M. Cespedes and J. Sun. Adaptive control of grid-connected inverters based on online grid impedance measurements. *IEEE Trans. on Sustainable Energy*, 5:516–523, 2014a.

M. Cespedes and J. Sun. Mitigation of inverter-grid harmonic resonance by narrow-band damping. *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2(4):1024–1031, 2014b.

A.P. Deshpande, B.N. Chaudhari, and V.N. Pande. Design and simulation of back-to-back converter for modern wind energy generation system using dSPACE. In *Proc. Internation conference on Power, Signals, Controls and Computation*, pages 1–6, 2012.

Z.A. Ghani, M.A Hannan, and A. Mohamed. Development of three-phase photovoltaic inverter using dSPACE DS1104 board. In *Proc. IEEE Student Conference on Research and Development*, pages 242–245, 2009.

K.R. Godfrey. *Perturbation Signals for System Identification*. Prentice Hall, UK, 1993.

W. Hu, H. Zhou, J. Sun, Y. Jiang, and X. Zha. Resonance analysis and suppression of system with multiple grid-connected inverters. In *Proc. IEEE International Future Energy Electronics Conference*, pages 1–6, 2015.

M. Lu, X. Wang, F. Blaabjerg, and C. Loh. An analysis method for harmonic resonance and stability of multi-paralleled lcl-filtered inverters. In *Proc. IEEE International Symposium on Power Electronics for Distributed Generation Systems*, pages 1-6, 2015.

A. Monti, E. Santi, R.A. Dougal, and M. Riva. Rapid prototyping of digital controls for power electronics. *IEEE Trans. on Power Electronics*, 18:915–923, 2003.

T. Roinila, M. Vilkko, and J. Sun. Broadband methods for online grid impedance measurement. In *Proc. IEEE Energy Conversion Congress and Exposition*, pages 3003–3010, 2013.

T. Roinila, M. Vilkko, and J. Sun. Online grid impedance measurement using discrete-interval binary sequence injection. *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2:985–993, 2014.

J. Sun. Impedance-based stability criterion for grid-connected inverters. *IEEE Trans. on Power Electronics*, 26:3075–3078, 2011.

D. Sundararajan. *The Discrete Fourier Transform: Theory, Algorithms and Applications*. World Scientific Publishing Co. Pte. Ltd., 2001.

X. Wang, F. Blaabjerg, and W. Wu. Modeling and analysis of harmonic stability in an AC power-electronics-based power system. *IEEE Transactions on Power Electronics*, 29:6421–6432, 2014.

A. Yazdani and R. Iravani. *Voltage-Sourced Converters in Power Systems*. John Wiley and Sons, inc.Hoboken, New Jersey, 2010.

Semi-Discrete Scheme for the Solution of Flow in River Tinnelva

Susantha Dissanayake Roshan Sharma Bernt Lie

Department of Electrical Engineering, IT, and Cybernetics, University College of Southeast Norway, Porsgrunn, Norway, {roshan.sharma, bernt.lie}@usn.no

Abstract

The Saint-Venant equation is a mathematical model which could be used to study water flow in an open channel, river, etc. The Kurganov-Petrova (KP) method, which is a second-order scheme, is used to solve the Saint-Venant equations with good stability. The water flow of a river between two hydropower stations in Norway has been simulated in this study using MATLAB and OpenModelica. The KP scheme has been used to discretize the Saint-Venant equations in the spatial domain, yielding a collection of Ordinary Differential Equations (ODEs). These are then integrated with time using the variable step-length solvers in MATLAB: ode23t, ode23s, ode45, and fixed step-length solvers: The Euler method, the second and fourth order Runge Kutta method (RK2 and RK4). In OpenModelica built-in, variable step-length DASSL solver has been used. From the simulation, it was observed that all solvers produce more or less similar results. Volumetric flowrate calculation indicated numerical oscillation with variable step-length solvers in MATLAB. The results indicated that it is reasonable to match the order of space and time discretization.

Keywords: semi-discrete KP scheme, OpenModelica, MATLAB

1 Introduction

By the year 2020, the 20-20-20 goal is to be achieved within the European Union: 20% efficiency in the improvement of power utilization, 20% reduction of carbon dioxide emission and 20% increment of renewable sources in the total energy mix (Blindheim, 2015). Subsequently, the utilization of renewable energy sources such as wind, hydro, and solar have to be optimized. Hydropower is a source of kinetic energy, which is extracted from flowing water. It is one of the mature renewable energy technologies in the current energy sector.

Norway is prominent in the production of hydropower as one of the renewable energy sources (Blindheim, 2015). Even though reservoir based power production technologies are well developed, power generation based on run-of-river systems are also common. As several hydropower stations are installed at different locations along the same river length, water flow between different hydropower stations influence their operations. When the upstream station (first station) increases its power production, volumetric flow of water out from the first station increases, thus the downstream power station (second station) has to increase the power production in order to utilize the water resource efficiently (Vytvytskyi et al., 2015).

Hence, it is vital to have an understanding of the propagation of the water flow from one station to the other, the change of water level at the second dam, the speed of the wave that hits the second dam, etc. Water flow modeling is also useful in other areas, e.g., managing water resources efficiently in agriculture, manage municipal drinking water distribution system and other applications in addition to power generation.

In this study, the flow of water in river Tinnelva in Southeast Norway between two hydropower stations are being considered. One power station is located at Årlifoss, and the other station is located at Grønvollfoss (downstream). The aim is to study the use of a semi-discrete scheme for the solution of flow in river Tinnelva. Objectives are to find an accurate and robust scheme for use in the control algorithm.

The paper is arranged as follows. The basic introduction to the governing equation and computational fluid dynamics (CFD) will be given in Section 2. Introduction to the KP numerical scheme will be provided in Section 3. Section 4 focuses on computer simulation. The Saint-Venant equation and the Kurganov-Petrova (KP) scheme as numerical scheme were used in simulation in order to compute final water level at Grønvollfoss dam. MATLAB and OpenModelica are being used as simulation software, and both built-in, variable step-length solvers and fixed step-length solvers are being used for the time integration. Parameters, assumptions, simplifications of the complex river system are also introduced in Section 4. Simulation results will be discussed in Section 5 together with numerical stability analysis.

2 Governing Equation for Flow Modeling

Conservation of properties of fluid flow, such as mass, energy, and momentum equations are important principles in fluid dynamics (Versteeg and Malalasekera, 2007). For the study of wave propagation, water flow, tsunami, etc. mathematical models have been derived based on the continuity equation and the momentum balance (Fayssal et al., 2015).

The Saint-Venant equation or commonly known as the 1-Dimensional (1D) shallow water equation, is used for

decades for simulation of water flow in open channels, rivers, etc. (Benkhaldoun et al., 2015). Basic conservation laws, such as momentum and mass conservation provide the base for the Saint-Venant equation which has been derived by integrating the momentum equation over the vertical coordinate (Benkhaldoun et al., 2015). This model provides stable solutions even at hydraulic jumps. The Saint-Venant equation can be posed as follows (Sharma, 2015).

$$\frac{\partial U}{\partial t}+\frac{\partial F}{\partial x}=S \tag{1}$$

where U vector is the vector of conserved variables

$$U=(A,Q)^T \tag{2}$$

F is the vector of fluxes

$$F=\left(Q,\frac{Q^2}{A}+gI_1\cos(\phi)\right)^T \tag{3}$$

and S is the source term

$$U=(z,q)^T, \tag{4}$$

$$F=\left(q,\frac{q^2}{z-B}+\frac{g}{2}(z-B)^2\right)^T, \tag{5}$$

$$S=\left(0,-g(z-B)\frac{\partial B}{\partial x}+\frac{gn^2q|q|(w+2(z-B))^{\frac{4}{3}}}{w^{\frac{4}{3}}}\frac{1}{(z-B)^{\frac{7}{3}}}\right). \tag{6}$$

Here, z is the water level above a datum, B is the bottom elevation from the datum, q is volumetric flow rate per unit width, w is the width of the river, n is Mannings roughness coefficient, and g is acceleration due to gravity. The S terms reflect source terms: including expressions of friction which give resistance against flow.

3 KP Numerical Scheme

In computational Fluid Dynamics (CFD), The Finite Volume Method (FVM) is based on averaging the Control Volume (CV) (Kurganov and Tadmor, 2000). As FVM average each CVs, discontinuities may occur at CV interfaces. This problem was recognized as the Riemann problem (Kurganov and Levy, 2002). In order to handle the Riemann problem, the Riemann solvers were developed (Kurganov and Tadmor, 2000). However, by the emerging of computer-based complex calculations, fast convergence with higher accuracy has to be accomplished. Subsequently, several novel techniques that could eliminate Riemann solvers were developed. The KP scheme was one of the developments which could handle discontinuities at CV interfaces without the Riemann solvers (Kurganov and Tadmor, 2000). The KP scheme is semi-discrete in nature: discretization in space and Ordinary Differential Equation (ODE) solvers in MATLAB and OpenModelica can be used to solve the resulting differential algebraic equations.

Kurganov and Petrova have developed a new scheme which could be considered as an extension/further development of the Nessyahu-Tadmor (NT) scheme (Kurganov and Tadmor, 2000). The NT scheme was developed to average the CV value by using the non-smooth Riemann fans over a fixed length $\triangle x$ (Nessyahu and Tadmor, 1990). In the KP scheme development, instead of averaging the non-smooth parts of the Riemann fans, precise local velocities of wave propagation have been considered along with small CVs of variable size (Kurganov and Tadmor, 2000). When the CV interface has discontinuities, a staggered CV concept can be introduced to eliminate the problem (LeVeque, 1999). During the transient, the local velocities are usually different at each side of a CV interface. Therefore, altered staggering at both sides of the CV is reasonable. Thus, the size of the virtual CVs are defined for a small time ($\triangle t$) by considering the local velocity of wave propagation. For each non-uniform CVs, a piecewise linear reconstruction has been done over the solution domain. Later the linear reconstructed values have been projected to the original uniform CVs while assuming the limits $\triangle t \to 0$ (Kurganov and Tadmor, 2000).

In the KP scheme, properties are indexed by a plus (+) and minus (-) with reference to the direction of the property flux. The local speed of discontinuity propagation has been calculated by considering the Jacobi matrix of the governing equations. In order to achieve higher resolution and a well-balanced scheme, the Total Variant Diminishing (TVD) concept together with the flux limiter concept has been used. The standard *minmod* limiter has been used in the original development of the KP scheme; many alternative flux limiters can be used just as well (Kurganov and Tadmor, 2000).

The KP scheme does not use the Riemann solvers. Hence, computational time can be reduced. Numerical viscosity with the KP scheme is lower compared to the NT scheme (Kurganov and Tadmor, 2000).

The KP scheme discretizes the Saint-Venant equation spatially, yielding a collection of ODEs in time. These ODEs can be written as follows,

$$\frac{d}{dt}\bar{u}_j=-\frac{H_{j+\frac{1}{2}}(t)-H_{j-\frac{1}{2}}(t)}{\triangle x}, \tag{7}$$

where $\bar{u}_j$ is the cell center average values, $H_{j\pm\frac{1}{2}}(t)$ are the central upwind numerical fluxes at the cell interfaces, defined as:

$$H_{j+\frac{1}{2}}(t)=\frac{a^+_{j+\frac{1}{2}}F\left(u^-_{j+\frac{1}{2}},B_{j+\frac{1}{2}}\right)-a^-_{j+\frac{1}{2}}F\left(u^+_{j+\frac{1}{2}},B_{j+\frac{1}{2}}\right)}{a^+_{j+\frac{1}{2}}-a^-_{j+\frac{1}{2}}}+\frac{a^+_{j+\frac{1}{2}}a^-_{j+\frac{1}{2}}}{a^+_{j+\frac{1}{2}}-a^-_{j+\frac{1}{2}}}\left(u^+_{j+\frac{1}{2}}-u^-_{j+\frac{1}{2}}\right), \tag{8}$$

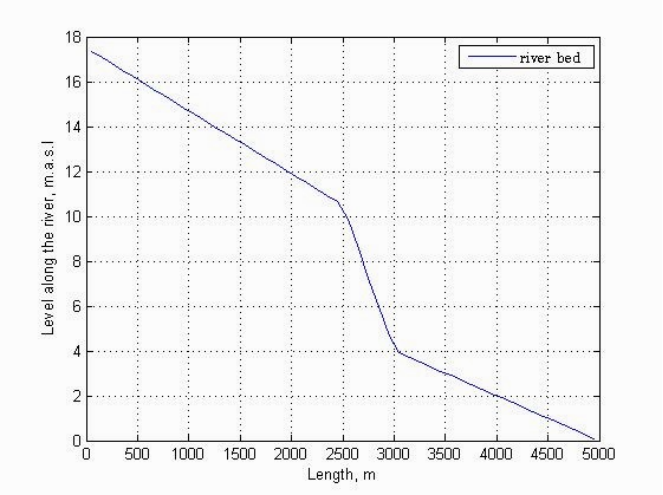


Figure 1. Bottom topography of the river.

$$H_{j-\frac{1}{2}}(t) = \frac{a^+_{j-\frac{1}{2}} F\left(u^-_{j-\frac{1}{2}}, B_{j-\frac{1}{2}}\right) - a^-_{j-\frac{1}{2}} F\left(u^+_{j-\frac{1}{2}}, B_{j-\frac{1}{2}}\right)}{a^+_{j-\frac{1}{2}} - a^-_{j-\frac{1}{2}}} + \frac{a^+_{j-\frac{1}{2}} a^-_{j-\frac{1}{2}}}{a^+_{j-\frac{1}{2}} - a^-_{j-\frac{1}{2}}}\left(u^+_{j-\frac{1}{2}} - u^-_{j-\frac{1}{2}}\right), \quad (9)$$

where $a^{\pm}_{j\pm\frac{1}{2}}$ are the one-sided local speeds of propagation and $u^{\pm}_{j\pm\frac{1}{2}}$ are property fluxes at indexed positions (Sharma 2015).

4 Simulation of the River Flow

In river Tinnelva, two hydropower stations, one at Årlifoss, and the other at Grønvollfoss are being operated by Skagerak Energi. The river reach is 5km in length. In the study, the bottom topography of the river has been divided into three sections of different slope, and the width of the river is assumed to be constant (180 m) during the whole reach of interest. The assumed bottom topography of the river is illustrated in Figure 1. The section from 2.5 km to 3 km has a steeper bed compared to the other sections (Vytvytskyi et al., 2015).

Due to different operational conditions, the volumetric outflow of water at the Årlifoss station is varying. The volumetric flow rates and other quantities are displayed in Table 1.

Other than the spatial discretization done by the KP scheme, time discretization methods in fixed step-length: The Euler method, the second order, and the fourth order Runge Kutta (RK2 and RK4) and the built-in variable step-length solvers: ode solvers in MATLAB (ode23s, ode23t, ode45) and the DASSL solver in OpenModelica, have been used. In addition to the water level computation, the numerical stability of each solver has been analyzed. Here, only for the numerical stability analysis, the variable step-length ode15s solver has been used.

Table 1. Results analysis.

Length	5 km
Number of CV	200
Time steps ($\triangle t$)	0.25 s
Volumetric flow in	120 m^3/s
Volumetric flow out	120 m^3/s
Volumetric flow increased	160 m^3/s
Width of the river	180 m
Initial water level at the dam	17.5 m
Mannings friction factor	0.04 $s/m^{1/3}$
Gravitational constant	9.81 m/s^2

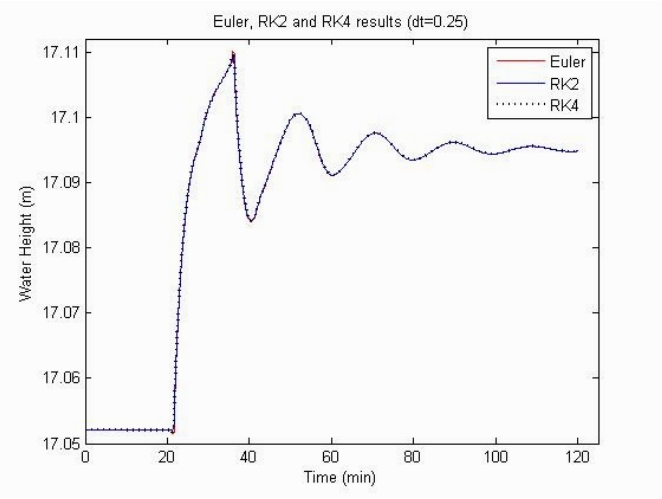


Figure 2. Fixed step-length solvers the Euler method, RK2, RK4 ($\triangle t = 0.25s$).

5 Results and Discussion

Results of the simulation study and numerical stability analysis will be discussed in the following sub sections.

5.1 Simulation Results

The built-in variable step-length solvers: ode23t and ode23s have second-order accuracy (Gladwell et al., 2003). Fixed step-length solvers: The Euler method, the RK2 method, and the RK4 method have first order, second order, and fourth order accuracy respectively (Gerald and Wheatley, 2004; LeVeque, 1992). ode45 and ode15s have higher order; higher than a second order of accuracy (Gladwell et al., 2003).

The simulation results using fixed step-length solvers (the Euler method, the RK2 method and the RK4 method) with fixed step length ($\triangle t$ = 0.25 second) show very similar behaviors as shown in Figure 2.

An exploded view of Figure 2 in the time range of 35 min to 37 min is shown in Figure 3. In the exploded view, the Euler method shows some deviation from the other two solutions. However, this deviation is very small.

Both RK2, RK4 and the Euler method algorithms show accuracy up to fourth significant digits in their final solu-

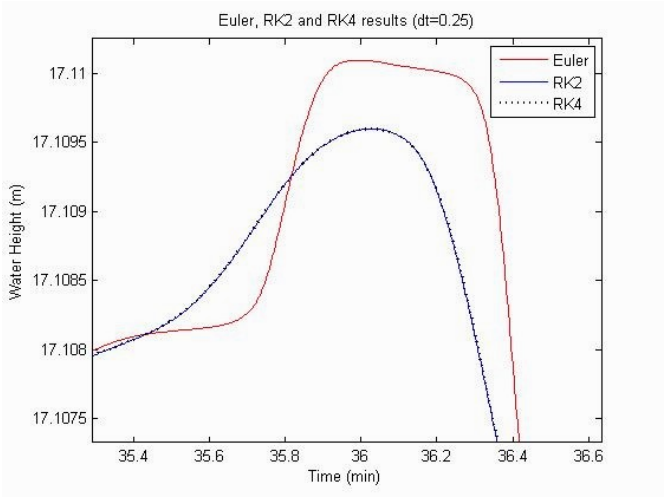


Figure 3. Exploded view of fixed step-length solvers (Euler method, RK2, RK4).

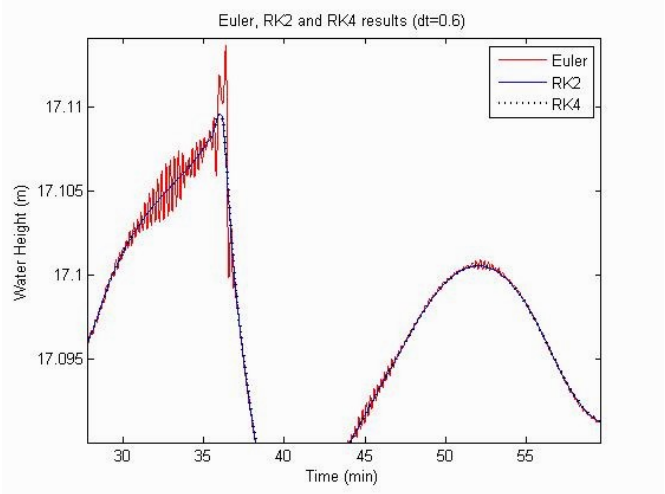


Figure 4. Fixed step-length solvers the Euler method, RK2, RK4 at $\triangle t$ = 0.6 s.

tions.

The solution obtained using the Euler method is highly dependent the choice of the step length; $\triangle t$. When $\triangle t$ is set to 0.6 seconds, the Euler method shows some oscillation in its solution (Figure 4).

With $\triangle t$ set to 0.7 seconds, the solution becomes unstable. While The Euler method shows higher oscillatoric behavior and unstable solutions when $\triangle t$ increases ($\triangle t \geq$ 0.7s) the RK2 and RK4 methods produce stable solutions. However, increment of $\triangle t$ has necessarily to be agreed with the CFL condition (Silvester et al., 2015; LeVeque, 1992) which is commonly written as,

$$C = \frac{u\triangle t}{\triangle x} \leq C_{max} \tag{10}$$

Here C is dimensionless number u refers the magnitude of the velocity, $\triangle x$ refers the length of CV. For the upwind scheme, $C_{max} = 1$ (Silvester et al., 2015).

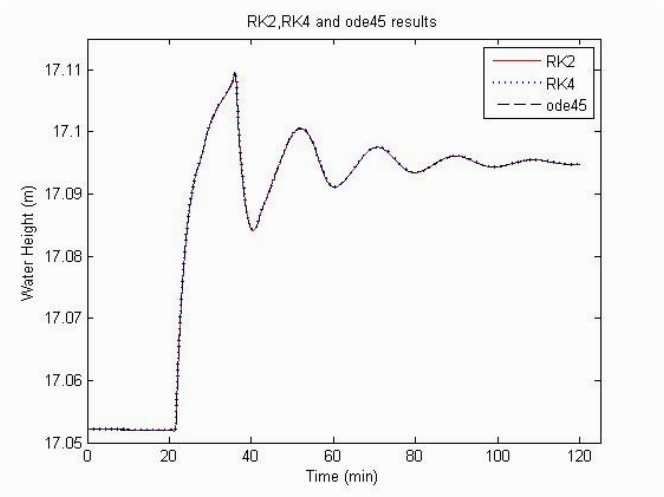


Figure 5. The RK2, RK4 method and ode45 solver results at $\triangle t$ = 0.25 s.

The standard KP scheme is a second order scheme in spatial discretization (Kurganov and Tadmor, 2000). Higher order time integrators: the RK4 method and ode45 were used to solve second order ODEs returned by spatial discretization of the Saint-Venant equation.

The idea was to check whether the higher order time integrators; higher than the second order, have a significant influence on solving second order ODEs more accurately.

According to the observations, both RK2 and the RK4 schemes show very similar solutions. Hence, this denotes that the higher order time integrators have the minor influence on the accuracy when it uses to solve lower order ODEs. The selection of an order of the time discretization that exceeds the order of the spatial discretization does not necessarily produce a more accurate solution (Liu and Tadmor, 1998). Consequently, in order to acquire a reasonably accurate solution, the order of the time discretization should be of either lower or the same order as the order of the spatial discretization.

When comparing variable step-length ode45 solvers in MATLAB with fixed step-length solvers: RK2 and RK4, all solvers produce very similar results (Figure 5). In exploded view (Figure 6), the variable step-length solver ode45 shows some minor oscillatory behavior. Even though the exploded view shows a small deviation, the results of the all three solvers (ode45, RK2, and RK4) compute the end-time water height at the Grønvollfoss dam accurately up to the fourth significant digit.

Results of all fixed step-length solvers: The Euler method, RK2, RK4 and all variable step-length solvers: ode23s, ode23t, ode45 are shown in Figure 7. The computed final water height of each different solver are similar (Figure 7), however, when it considers closely zooming in, it is possible to see minor deviations.

OpenModelica simulation by using the built-in DASSL solver shows a similar pattern compared to the solutions of the MATLAB solvers as shown in Figure 8. Table 2 sum-

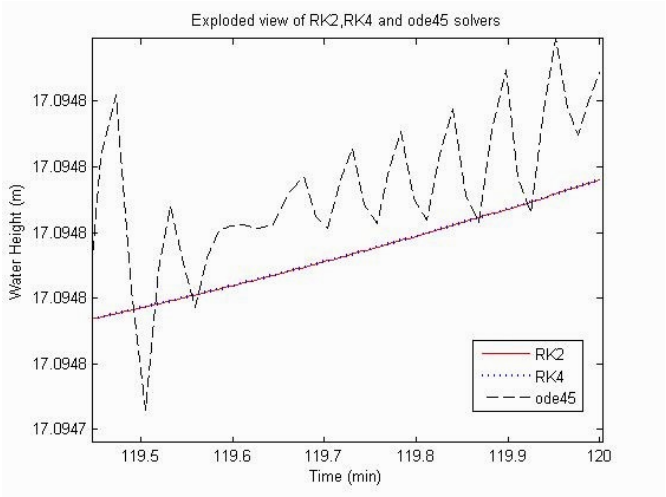


Figure 6. Exploded view of the RK2, RK4 method and ode45 solver results at $\triangle t = 0.25$ s.

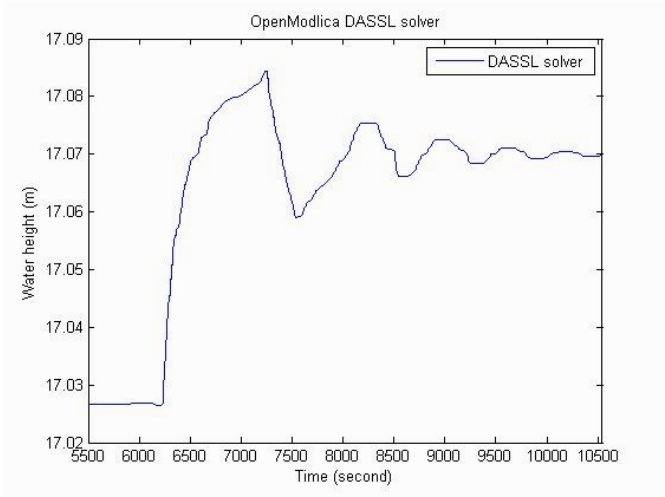


Figure 8. OpenModelica simulation results with DASSL solver.

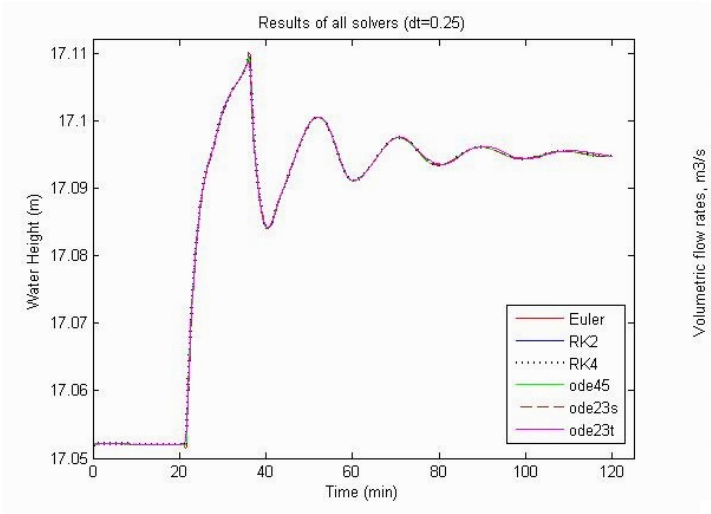


Figure 7. All solvers result at $\triangle t = 0.25$ s.

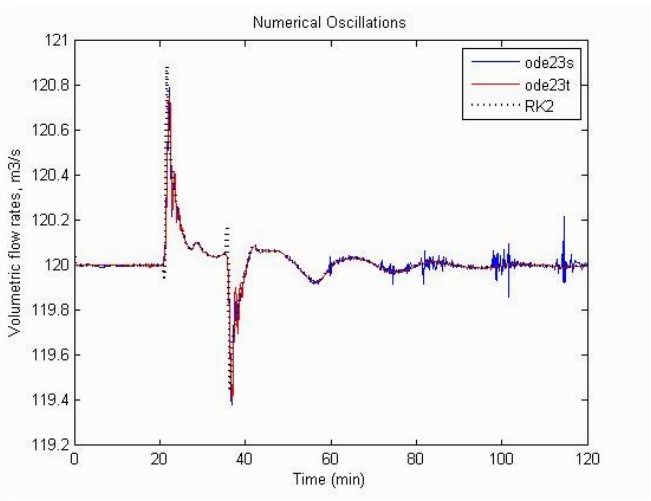


Figure 9. Numerical oscillation of group 01 solvers $\triangle t = 0.25$ s.

marizes other observations of the simulation study. Time consumed by each solver, minimum and maximum time step of variable step-length solvers and steady water level for all solvers are tabulated in the Table 2.

5.2 Simulation Results for Numerical Stability Analysis

In this section, results of numerical stability analysis will be discussed. For ease of comparison, the six solvers, which were used, have been divided into two groups based on their order of the accuracy. Thus, ode23s, ode23t and the RK2 method were classed into group 01, which are of second order in accuracy. The RK4, ode45 and ode15s solvers were classed into the group 02, which are of higher order in accuracy in time discretization than group 01.

The results of the volumetric flow rate calculation for the lower order group (group 01) are plotted in Figure 9. From the observations, the ode23s and ode23t solvers show higher oscillation in volumetric flow rate calculations than RK2.

For the results of the group 02 solvers, Figure 10, it can be clearly seen, that the oscillatory nature increases with increasing order of the time discretization. The solution using variable step-length ode solvers are more oscillatory compared to the fixed step-length solvers RK4 for the volumetric flowrate calculations. ode45 shows higher oscillation while ode15s show relatively smaller oscillations for the volumetric flow rate calculation.

As a whole, it could be observed that built-in variable step-length ode solvers show a relatively oscillatory nature for the volumetric flow rate calculations.

6 Conclusions

Based on the simulated results, it can be concluded that a higher order in the time discretization than the order in the spatial discretization does not necessarily produce more accurate solutions, consequently, matching orders of both

Table 2. Results analysis.

Description	*Solver*	*Values*
Computational time at $\triangle t = 0.25$ in seconds	ode23t (variable step-length)	11
	ode23s (variable step-length)	325
	ode45 (variable step-length)	15
	ode15s (variable step-length)	125
	The Euler method (fixed step-length)	29
	RK2(fixed step-length)	52
	RK4(fixed step-length)	105.353
[Min, max] time steps for ode solvers	ode23t	[1.022,196]
	ode23s	[3.2076,133]
	ode45	[0.5533,1.37]
	ode15s	[0.6381,150]
Steady water level in front of Grønvollfoss dam	For all solvers	17.0948 (m)

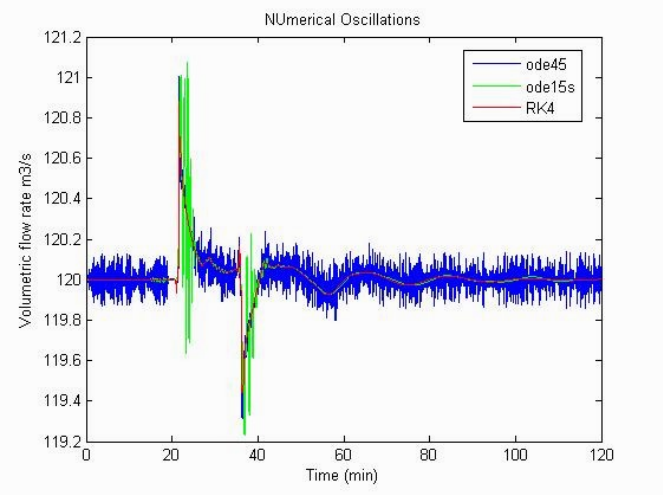


Figure 10. Numerical oscillation of group 02 solvers $\triangle t = 0.25$ s.

spatial and time discretization is a good idea. Choice of $\triangle t$ necessarily has to be agreed with the CFL condition in order to achieve convergence with satisfactory accuracy in the final solution. For a selected $\triangle t$, which is higher than 0.7s, the Euler method produces oscillatory solution apart from the chosen $\triangle t$ satisfies the CFL condition. However, the RK2 and RK4 methods are quite stable while the Euler method shows oscillations. The numerical stability analysis indicated that the higher order variable step-length solvers are more oscillatory compared to higher order fixed step-length solvers. Final water height at Grønvollfoss dam is more or less similar with compared to different computations with variable step-length solvers and fixed step-length solvers. Results of the simulation study highly depend on the assumption made prior to simulation. The studied KP scheme has been found to be efficient and robust, and in a form suitable for use in control algorithm.

Acknowledgements

Kindly convey sincere thanks to the other project group members: Janitha Chandimal, Junyang Mao, and Obianuju Ezuka.

References

Fayssal Benkhaldoun, Saida Sari, and Mohammed Seaid. Projection finite volume method for shallow water flows. *Mathematics and Computers in Simulation*, 118:87 – 101, 2015. ISSN 0378-4754. doi:10.1016/j.matcom.2014.11.027.

Bernt Blindheim. A missing link? the case of norway and sweden: Does increased renewable energy production impact domestic greenhouse gas emissions? *Energy Policy*, 77:207 – 215, 2015. ISSN 0301-4215. doi:10.1016/j.enpol.2014.10.019.

Curtis F. Gerald and Patrick O. Wheatley. *Applied Numerical Analysis*. Pearson Education Inc., Boston, 2004.

Ian Gladwell, Larry Shampine, and S. Thompson. *Solving ODEs with MATLAB*. Cambridge University Press, New York, NY, USA, 2003. ISBN 0521530946.

Alexander Kurganov and Doron Levy. Central-upwind schemes for the Saint-Venant system. *Mathematical Modelling and Numerical Analysis*, 36(3): 397–425, 2002. doi:10.1051/m2an:2002019.

Alexander Kurganov and Eitan Tadmor. New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations. *Journal of Computational Physics*, 160(1):241 – 282, 2000. ISSN 0021-9991. doi:10.1006/jcph.2000.6459.

Randall J. LeVeque. *Numerical Methods for Conservation Laws*. Lectures in Mathematics, ETH Zürich. Birkhäuser Basel, 2nd edition, 1992. ISBN 978-3-0348-8629-1. doi:10.1007/978-3-0348-8629-1.

Xu-Dong Liu and Eitan Tadmor. Third order nonoscillatory central scheme for hyperbolic conservation laws. *Numerische Mathematik*, 79(3):397–425, May 1998. ISSN 0945-3245. doi:10.1007/s002110050345.

Haim Nessyahu and Eitan Tadmor. Non-oscillatory central differencing for hyperbolic conservation laws. *Journal of Computational Physics*, 87(2):408–463, April 1990. ISSN 0021-9991. doi:10.1016/0021-9991(90)90260-8.

Roshan Sharma. Second order scheme for open channel flow. Technical report, USN Open Archive, University College of Southeast Norway, 2015. URL `http://hdl.handle.net/11250/2438453`.

David J. Silvester, John W. Dold, and David Francis Griffiths. *Essential Partial Differential Equations: Analytical and Computational Aspects*. Springer International Publishing, 2015. ISBN 978-3-319-22569-2.

H. K. Versteeg and W. Malalasekera. *An introduction to computational fluid dynamics*. Pearson Education, Upper Saddle River, United States, 2nd edition, 2007. ISBN 9780131274983.

Liubomyr Vytvytskyi, Roshan Sharma, and Bernt Lie. Model based control for run-of-river system. Part 1: Model implementation and tuning. *Modeling, Identification and Control*, 36(4):237–249, 2015. doi:10.4173/mic.2015.4.4.

Simulation of Glycol Processes for CO_2 Dehydration

Lars Erik Øi Birendra Rai

Department of and Process, Energy and Environmental Technology, University College of Southeast Norway, Norway

lars.oi@usn.no

Abstract

Water must be removed from CO_2 prior to transport or storage to avoid corrosion and hydrate formation. Absorption into triethylene glycol (TEG) followed by desorption is the traditional gas dehydration method, and is expected to be the preferred method for large scale CO_2 dehydration. There is no agreement on the level of accepted water content after dehydration, and the specifications vary normally in the range between 5 and 500 ppm (parts per million by volume). In literature, it is claimed that use of solid adsorbents is necessary to reduce the water content to below 30 or 10 ppm. In this work, simulations in Aspen HYSYS demonstrate that it is possible to obtain below 1 ppm water using a traditional glycol dehydration process including an extra stripping column. The models Peng-Robinson (PR) and Twu-Sim-Tassone (TST) with updated parameters in Aspen HYSYS version 8.0 are used. A Drizo process using a stripping gas which is later condensed and recirculated is also simulated, and this process also achieves a water content below 1 ppm in dehydrated CO_2.

Keywords: CO_2, glycol, dehydration, Aspen HYSYS

1 Introduction

CO_2 removed from natural gas or from CO_2 capture should be dehydrated prior to transport or storage. Water may lead to problems like corrosion and hydrate formation. The need for water removal from CO_2 and possible specifications are discussed in several references (Cole et al., 2011; Uilhorn, 2013; Buit, 2011) and water specifications are normally in the range between 5 and 500 ppm (parts per million by volume). CO_2 for enhanced oil recovery normally requires the lowest water content.

There are several different gas dehydration methods available. The most used processes for dehydration are based on absorption, adsorption or membranes. The most traditional method for large scale dehydration to moderate water levels is by absorption into triethylene glycol (TEG). For very low water levels, adsorption processes are claimed to be necessary (Kohl and Nielsen, 1997; Kemper et al., 2014). There are however simulated reasonable processes for glycol dehydration of CO_2 down to water levels below 5 ppm (Øi and Fazlagic, 2014) using stripping gas and an extra stripping column. The Drizo process makes use of a condensable stripping gas which is recirculated and is able to reduce the water content down to 1 ppm (Prosernat, 2016).

A recent study (Kemper et al., 2014) has evaluated different commercial processes for CO_2 dehydration based partially on information from vendors of technology. Processes based on glycol and on solid adsorption were evaluated. It was claimed that use of solid adsorbents is necessary to reduce the water content to below 30 or 10 ppm.

The main purpose of this paper is to present updated simulations of flow-sheet models for CO_2 dehydration by absorption in triethylene glycol. The Peng-Robinson (PR) model and the Twu-Sim-Tassone (TST) model (the glycol package in Aspen HYSYS) are used. In version 8.0, the parameters in the TST model have been updated compared to version 7.2 which was used in earlier work with Aspen HYSYS (Øi and Fazlagic, 2014). Especially, it is an aim to simulate the alternative including an extra stripping column and a Drizo process achieving a water level down to 1 ppm. The simulation results in this work are mainly from a Master Thesis work (Rai, 2016).

2 Simulation Programs and Models for Glycol Dehydration

Commercial process simulation programs which have been used for glycol dehydration are Aspen Plus, Aspen HYSYS, Pro/II and ProMax. Process simulation programs are useful for simulation of absorption processes because complex vapour/liquid equilibrium models are available in the programs and because efficient stage to stage models for absorption and desorption columns are available.

Absorption and distillation columns are traditionally simulated as a sequence of equilibrium stages. The stages can also be specified with a Murphree efficiency. A Murphree efficiency can be defined as the change in mole fraction of a component (in this case water) from the stage below to a given stage, divided by the change if equilibrium was achieved on the given stage (Kohl and Nielsen).

Vapour/liquid equilibrium data in the TEG/water system has been discussed in several papers (Kohl and Nielsen, 1997; Øi and Fazlagic, 2014; Øi, 1999). The vapour/liquid equilibrium between CO_2 and water has also been evaluated (Cole et al., 2011; Austegard, 2006). This equilibrium must be included in a complete model of the TEG/water/CO_2 system. One special challenge for such a model, is to calculate the correct solubility of CO_2 in a TEG/water solution.

In Aspen HYSYS, the equilibrium models PR (Peng and Robinson, 1976) and TST (Twu et al., 2005) are available for glycol dehydration. The TST model is claimed in the Aspen HYSYS program documentation to be the most accurate. This is however based on the assumption of dehydration of natural gas, and it is uncertain whether the TST model is accurate when CO_2 is the dominating gas. The TST model parameters have been updated from earlier versions in the Aspen HYSYS version 8.0 used in this work. The PR model has only one adjustable parameter for each binary component pair while TST has 5 adjustable parameters for each binary pair. In Aspen Plus, PRO/II and ProMax there are also models available recommended for the TEG/water system (Øi and Fazlagic, 2014).

Simulations of the natural gas dehydration process in Aspen HYSYS have been performed with emphasis on glycol regeneration by Øi and Selstø (2002). A traditional TEG dehydration process was simulated using the PR equation of state. A water content of 58 ppm was specified in the dehydrated gas. Different regeneration processes were simulated, e.g. addition of stripping gas to the reboiler and adding stripping gas to an extra stripping column. Bilsbak (2009) used the simulation program Pro/II to simulate TEG dehydration of CO_2 using an equation of state.

Besides the references mentioned in this work, there is very little information found in the open literature about simulation of CO_2 dehydration processes and particularly of CO_2 dehydration based on absorption in glycol.

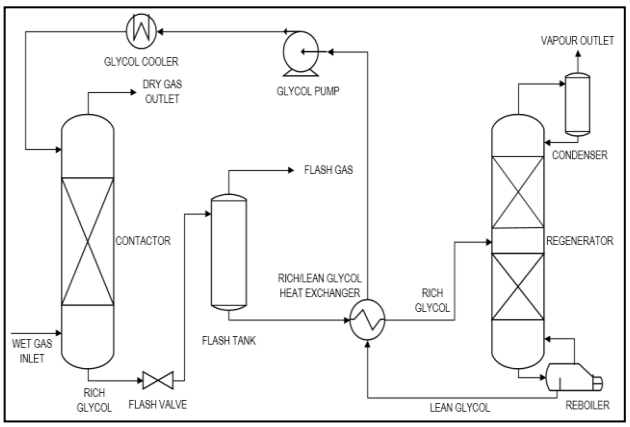


Figure 1. A traditional TEG dehydration process

3 Process Description

A traditional glycol dehydration process is shown in Figure 1. Water is absorbed from a gas into a glycol solution in an absorption column (contactor). The liquid (rich glycol) is then depressurized to a flash tank to evaporate some of the absorbed CO_2. Then the liquid is heated by regenerated glycol in a heat exchanger and sent to a desorber (regenerator). In this column, water is removed from the top and regenerated TEG is removed from the bottom. Heat is added in the reboiler. The regenerated TEG (lean glycol) is pumped through the heat exchanger and a cooler back to the absorber. More detailed process descriptions can be found in e.g. (Kohl and Nielsen, 1997; Øi and Fazlagic, 2014). It is possible to reduce the water content in the glycol by adding stripping gas to the reboiler. The reduction in water content in the glycol will make it possible to improve the dehydration of the CO_2 in the treated gas. The stripping gas can come from the flash gas or from the dehydrated gas. The water content in the glycol can be reduced further by adding an extra stripping column below the reboiler. This is shown in Figure 2.

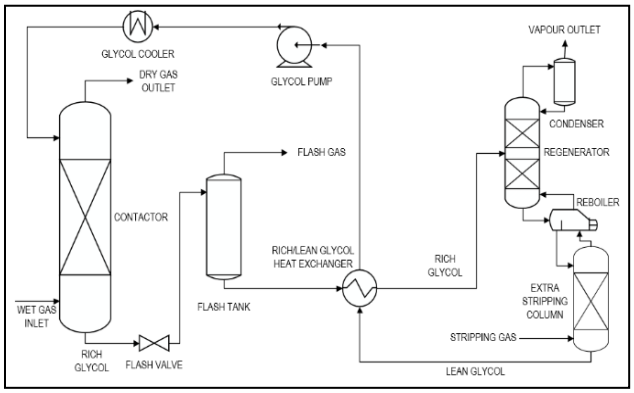


Figure 2. A TEG dehydration process with extra stripping column

There are also different special glycol dehydration processes. One example is the Drizo process (Prosernat, 2016) where the added stripping gas is a condensable component which is condensed after the stripping column and recirculated to the bottom of the stripping column. In earlier work (Øi and Fazlagic, 2014) the possible water content achievable from these processes was calculated using Aspen HYSYS. A traditional process could achieve below 200 ppm. Using stripping gas, less than 50 ppm could be achieved. Adding an extra stripping column below the reboiler could reduce the water content to below 5 ppm.

To achieve very low water content, the height of the absorption column has to be increased compared to a traditional process. The absorption column will contain plates or packing. An expected plate efficiency is order

of magnitude 50 %, and an expected packing efficiency of a packing height of 0.5 meter is also order of magnitude 50 % (Øi, 2003). An increase of the packing height from 10 to 20 stages or order of magnitude from 5 to 10 meters will increase the cost significantly. The extra stripping column will add complexity to the process. The cost for extra height in this column is however small because the diameter is very low.

4 Process Simulation, Results and Discussion

4.1 Simulation of Standard Process

A traditional TEG dehydration process as in Figure 1 has been simulated in the simulation program Aspen HYSYS version 8.0 using the PR equation of state and TST model (the glycol property package). The Aspen HYSYS flow-sheet model for the base case simulation is presented in Figure 4. The specifications for the base case process calculation are given in Table 1. These are similar to the specifications in earlier simulations (Øi and Fazlagic, 2014). The enthalpy setting in Aspen HYSYS was changed to the Cavett model instead of the default property package EOS (equation of state) to avoid unrealistic low temperatures.

Table 1. Specifications for base case simulation

Parameter	Value
Inlet gas temperature	30 °C
Inlet gas pressure	3000 kPa
Inlet gas molar flow rate	501.1 kmol/h
Water in inlet gas	0.23 mol-%
TEG to contactor temperature	35 °C
TEG to contactor pressure	3000 kPa
TEG to contactor, flow (in first iteration)	3.583 kmol/h
Water in lean TEG (in first iteration)	1.04 mass-%
Number of stages in absorber	10
Murphree efficiency in absorber stages	0.5
Pressure after depressurization valve	110 kPa
Temperature in TEG to regeneration	153 °C
Number of stages in desorber	4
Murphree efficiency in desorber stages	1.0
Reflux ratio in desorber	0.5
Reboiler temperature	200 °C
Desorber pressure	101 kPa
Pressure after TEG pump	3000 kPa

Recommendations for a traditional process can be found also in Kohl and Nielsen (1997). The absorption column was simulated with 10 stages and with Murphree efficiency 0.5 on each stage, which is assumed to be equivalent to approximately 10 actual plates or 5 meter of structured packing (Shresta, 2015).

The calculation sequence of the process in the Aspen HYSYS flow-sheet is mostly following the real flow direction. The gas feed stream to the absorber is saturated with water. The liquid feed to the absorber has to be estimated in the first iteration. Then the absorption column and the rest of the process is calculated step by step. The cold side of the main heat exchanger is calculated based on a specified temperature on the stream to the desorber.

Using the base case specifications in Table 1, the water content in dehydrated gas was calculated to 153 ppm with the PR model and 133 ppm with the glycol package using the TST model. Compared to earlier simulations, the results from the PR model were identical. The results using the TST model were lower in water content compared to the earlier results (Øi and Fazlagic, 2014). The differences are expected to be due to updated TST parameters. The deviations between calculated water content from PR and TST are lower in this work. The process was simulated also at other absorption pressures. The results are shown in Figure 3.

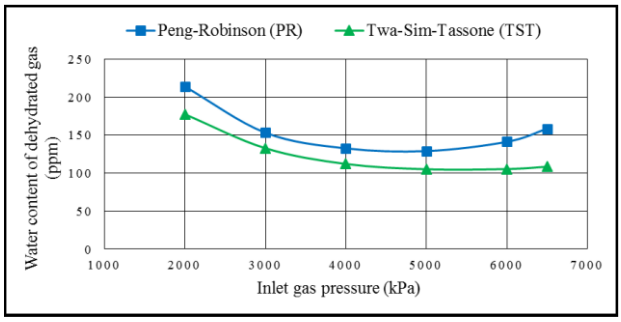


Figure 3. Water content in dehydrated gas as a function of absorption pressure

Both models gave a minimum water content and maximum dehydration efficiency at 5000 kPa. Minimum water content was 129 ppm at 5000 kPa using the PR model and 105 ppm using the TST model. In earlier work Øi and Fazlagic (2014) the TST model gave a minimum water content at 3000 kPa.

The dehydration can be improved by increasing the TEG circulating rate and by increasing the number of absorption stages in the standard process. However, these changes gave only minor improvements in achieved water content in the dehydrated gas.

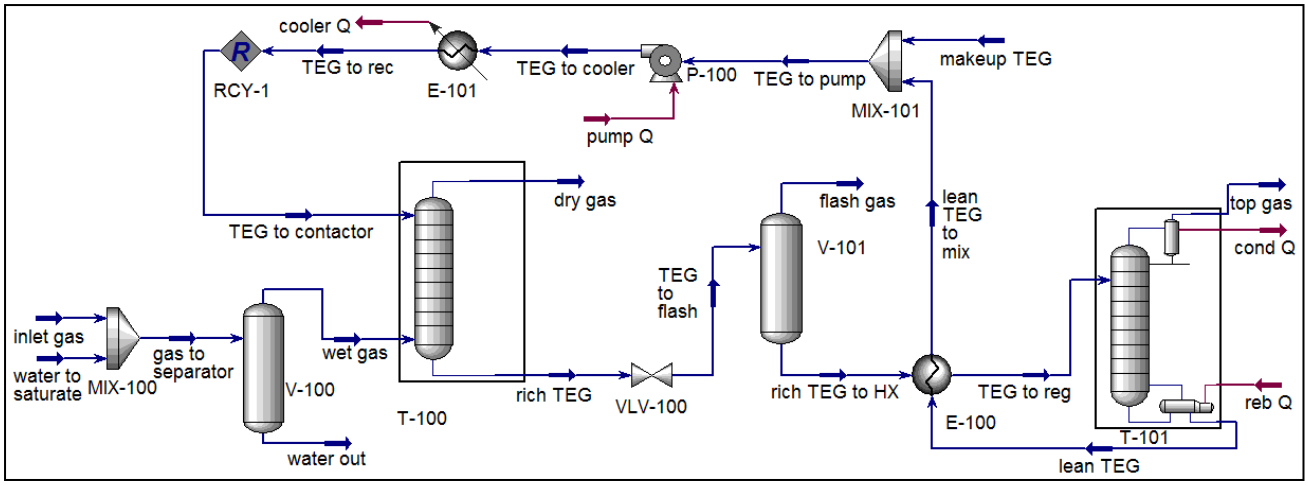


Figure 4. Aspen HYSYS flow-sheet model for traditional TEG dehydration process

4.2 Simulation of Stripping Gas to Reboiler

A process with stripping gas added to the reboiler was simulated. The stripping gas was specified with temperature 190 °C, pressure 101 kPa, 99.17 % CO_2 and 0.83 % water (Øi and Fazlagic, 2014). The composition was similar to the flash gas composition from depressurization after the absorption column in the base case calculation. Figure 5 shows the results from the Aspen HYSYS simulations using the TST model.

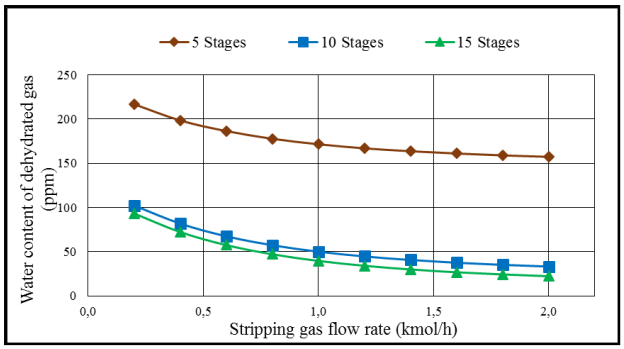


Figure 5. Water content as a function of stripping gas and number of absorber stages with stripping gas to reboiler using the TST model

Calculations comparable to the base case simulations were performed with varying the stripping gas flow and number of stages in the absorption column. Using the TST model, less than 50 ppm was achieved using 10 absorption stages and 1 kmole/h stripping gas. Using 15 stages, it was necessary to use 0.6 kmole/h to achieve less than 50 ppm in the dehydrated gas.

There is some difference in calculated flash gas amount for the two models, 0.89 kmole/h for the PR model and 0.63 kmole/h for the TST model. This indicates that the two models calculate the CO_2 solubility in the TEG/water solution differently. When using the PR model, the water amount for a given stripping gas amount is slightly higher than using the TST model. But because the PR model calculates a higher flash gas amount, the PR model also calculates about 50 ppm when the amount of stripping gas is set to the amount of available flash gas.

Increasing the number of stages and the amount of stripping gas made it possible to improve the dehydration down to about 30 ppm. This was achieved with both the PR and the TST model.

4.3 Simulation of Stripping Gas to Extra Column

A simulation of the process in Figure 2 was performed. In earlier work (Øi and Fazlagic, 2014) this process has been simulated using two strategies. The first strategy was to simulate the process with two columns. The second strategy was to simplify the flow-sheet model by simulating the desorption column and the extra stripping column as one column with heating at an intermediate stage as shown in Figure 6. Similar numerical results have been achieved using these two strategies (Øi and Fazlagic, 2014; Øi and Selstø, 2002). The strategy with only one column was chosen in this work because these simulations are easier to converge.

The extra stripping column was specified with 3 equilibrium stages. In the case of the extra stripping gas and the desorber simulated as one column, it was 3 stages between the stripping gas feed to the bottom and the reboiler heat addition. The number of stages above and below the feed in the desorber and in the extra stripping column have been varied in earlier work (Shresta, 2015). The results showed that increasing the number of stages in the sections above 3 (as specified in this work) did not improve the dehydration efficiency significantly.

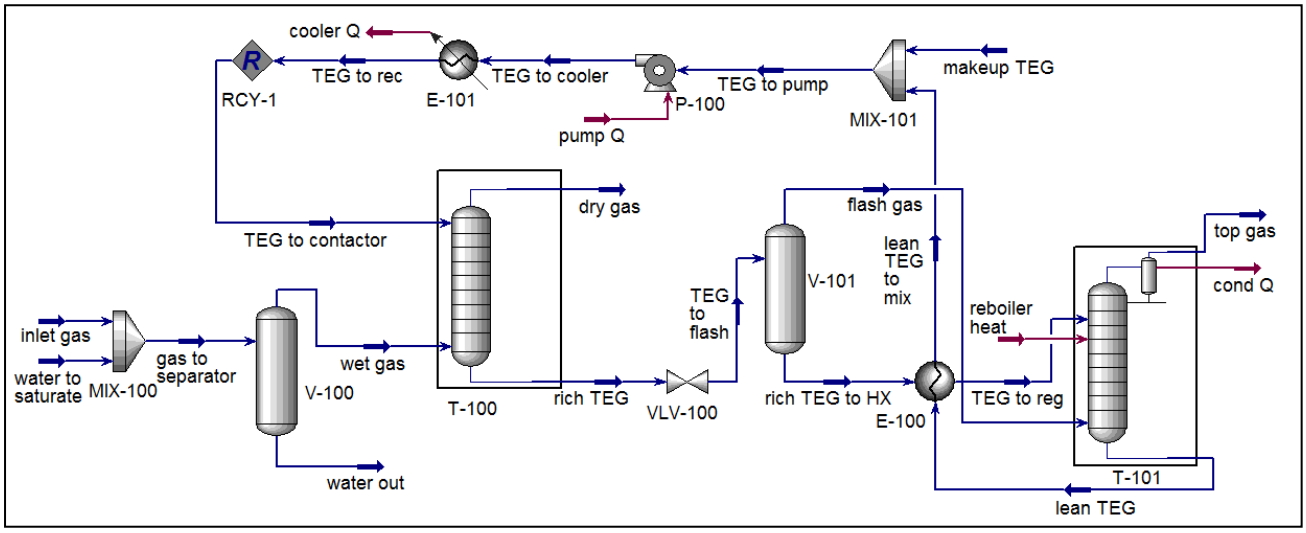


Figure 6. HYSYS flow-sheet of dehydration process with extra stripping column

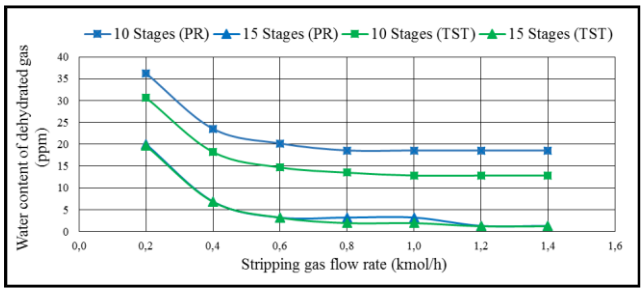


Figure 7. Water content as a function of stripping gas and number of absorber stages with extra stripping column

Figure 7 shows results from these calculations as a function of stripping gas and number of absorber stages for the process with an extra stripping column.

Using 10 absorber stages reduces the water content down to about 20 ppm using the PR model and down to about 15 ppm using the TST model. The water content is not reduced further by adding more stripping gas with 10 absorber stages. With 15 stages or more, less than 5 ppm was achieved using the available flash gas, and less than 2 ppm was achieved by adding more stripping gas. Using more than 15 absorber stages did not improve the dehydration efficiency significantly.

In earlier work (Øi and Fazlagic, 2014) similar results were obtained using 20 absorber stages for the same conditions. With more than 1.2 kmole/h, less than 5 ppm was achieved with both the PR model and the TST model.

At very low water content (below 2 to 5 ppm) the water amount in the stripping gas is a limiting factor. Less than 1 ppm water in dehydrated gas has been calculated when pure (dehydrated) CO_2 was used as stripping gas. This was calculated using 15 absorber stages and a high stripping gas amount with both the PR and the TST model.

4.4 Simulation of a Drizo process

A Drizo process with n-heptane as stripping gas was simulated. The flow-sheet is shown in Figure 8. The condensing of the stripping gas is performed in a three phase separator, and the organic phase is pumped back and evaporated in a heater before it is added to the bottom of the extra stripping column. It was checked that the amount of recirculated n-heptane was approximately equal to the n-heptane added to the desorber in the simulation.

A similar process has been simulated earlier (Øi and Selstø, 2002). An argument for not closing the loop in the simulation of the recirculation of the stripping gas, is that a small fraction of the added stripping gas will be dissolved in the solvent and lost e.g. in the flash gas. This is of minor interest for the evaluation of the process.

The results for n-heptane as a stripping gas using the TST model are presented in Figure 9. Similar results were obtained using the PR model. Similar simulations were also performed with the components benzene and toluene as stripping gas. When using benzene and toluene, more of the stripping gas was dissolved in the glycol, so that the deviation between added stripping gas and recirculated stripping gas became larger.

Similar levels of water were achieved as in the simulations with flash gas or CO_2 as stripping gas. Less than 20 ppm was obtained using 10 absorption stages in the absorber and available flash gas. Less than 5 ppm was obtained with 15 stages and stripping gas amounts similar to the available flash gas and less than 1 ppm when the amount of stripping gas was increased.

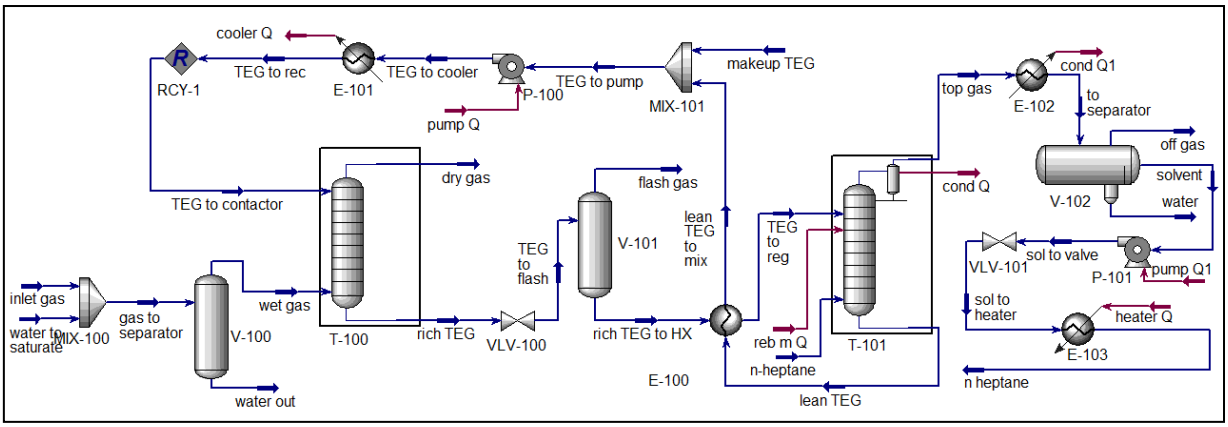


Figure 8. Aspen HYSYS flow-sheet of Drizo process using n-heptane as stripping gas

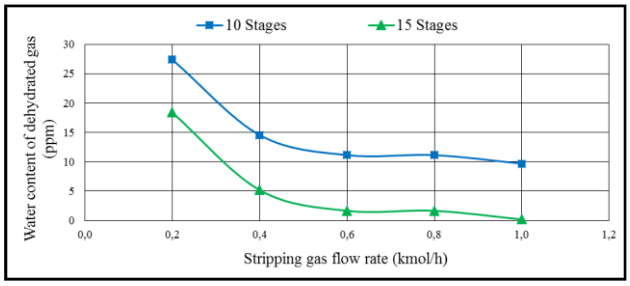


Figure 9. Water content in Drizo process with n-heptane as a function of stripping gas and number of absorber stages using the TST equilibrium model

5 Conclusions

It is demonstrated that it is possible to simulate both a traditional glycol dehydration process and more advanced CO_2 dehydration processes like the Drizo process using the process simulation program Aspen HYSYS. The PR and TST models give similar results for all the simulated alternatives. The calculated water content for different absorption pressures shows that a maximum dehydration efficiency is achieved at a pressure of about 5000 kPa.

A traditional TEG dehydration process is satisfactory to achieve a water content below 150 ppm in dehydrated CO_2. Using stripping gas, a water specification of less than 50 ppm can be achieved. In a Drizo process or a process including an extra stripping column and a high absorption column, it is possible to achieve less than 1 ppm water in dehydrated CO_2.

References

A. Austegard, E. Solbraa, G. de Koeijer and M.J. Mølnvik. Thermodynamic models for calculating mutual solubilities in H_2O-CO_2-CH_4 mixtures. *Trans IChemE, Part A, Chem. Eng. Res. Des.*, 84(A9):781-7946, 2011.

V. Bilsbak. *Conditioning of CO_2 coming from a CO_2 capture process for transport and storage purposes*. Master Thesis, NTNU, Trondheim, Norway, 2009.

L. Buit, M. Ahmad, W. Mallon and F. Hage. CO_2 EuroPipe study of the occurrence of free water in dense phase CO_2 transport. *Energy Procedia*, 4:3056-3062, 2011.

I.S. Cole, P. Corrigan, S. Sim and N. Birbilis. Corrosion of pipelines used for CO_2 transport in CCS: Is it a real problem? *International Journal of Greenhouse Gas Control*, 5(7):749-756, 2011.

J. Kemper, L. Sutherland, J. Watt and S. Santos. Evaluation and analysis of the performance of dehydration units for CO_2 capture. *Energy Procedia*, 63:7568-7584, 2014.

A.L. Kohl and R. Nielsen. *Gas purification, 5th ed.*, Gulf Publication, Houston. 1997.

D. Peng and D.B. Robinson. A New Two-Constant Equation of State. *Ind. Eng. Chem. Fundam.*, 15(1):59-646, 1976.

Prosernat. Glycol dehydration best process. http://www.prosernat.com/en/solutions/upstream/gas-dehydration/drizo.html, uploaded 11 April, 2016.

B. Rai. *CO_2 dehydration after CO_2 capture.* Master Thesis, University College of Southeast Norway, 2016.

S. Shresta. *Simulation of CO_2 dehydration after CO_2 capture*. Master Thesis, Telemark University College, Porsgrunn, Norway, 2015.

C.H. Twu, V. Tassone, W.D. Sim and S. Watanasiri. Advanced equation of state method for modeling TEG–water for glycol gas dehydration, *Fluid Phase Equilibria*, 228-229: 213-221, 2005.

F.E. Uilhorn. Evaluating the risk of hydrate formation in CO_2 pipelines under transient operation. *International Journal of Greenhouse Gas Control*, 14(5):177-182, 2013.

L.E. Øi. Calculation of dehydration absorbers based on improved phase equilibrium data. *78th Annual Convention of Gas Processors Association*, pp. 32-37, 1999.

L.E. Øi. Estimation of tray efficiency in dehydration absorbers. *Chem. Eng. Proc.*, 42(11):867-878, 2003.

L.E. Øi and M. Fazlagic. Glycol dehydration of captured carbon dioxide using Aspen Hysys simulation. *Linköping Electronic Conference Proceedings*, 2014.

L.E. Øi and E.T. Selstø. Process Simulation of Glycol Regeneration. *GPA Europe's meeting*, Bergen, Norway, 2003.

Mixing and Segregation of two Particulate Solids in the Transverse Plane of a Rotary Kiln

Sumudu Karunarathne[1] Chameera Jayarathna[2] Lars-Andre Tokheim[1]
[1]Department of Process, Energy and Environmental Technology, University College of Southeast Norway, Norway
{sumudu.karunarathne,lars.a.tokheim}@usn.no
[2]Tel-Tek, Norway, chameera.jayarathna@tel-tek.no

Abstract

Mixing of two granular phases in a rotary kiln was investigated through CFD simulations using a two-dimensional transverse plane based on the Eulerian approach and the kinetic theory of granular flows. Simulations were performed transverse with the aim to investigate mixing of two particulate solids, $CaCO_3$ and Al_2O_3, under the rolling mode. Simulation results indicated particle segregation rather than mixing during the plane rotation. Volume fractions and velocity contours of each phase were examined to understand the mixing and segregation. Particles with lower density and small particle diameter are collected in the middle section of the bed, while particles with a higher density and larger particle diameter get collected at the bottom of the rotating cylinder. Variations in densities and particle sizes of solid particles were identified as the main causes of the particle segregation. Further studies are required to examine the effect of degree of filling on mixing performance and how the use of lifters may improve the mixing efficiency.

Keywords: rotary kiln, granular flow, rolling mode, active layer, passive layer

1 Introduction

In industry, materials are needed to be processed in various ways to gain the desired quality of the product. Rotary kilns are widely accepted for the pyroprocessing of many types of materials in different industries owing to efficient mixing and heat transfer performances (Liu et al., 2016). Understanding of particles mixing inside the kiln is vital to enhance heat transfer performance within the bed that improvise material conversion rates in pyro processing.

Granular flows of a transverse plane in a rotary kiln can be categorized into six transport modes; slipping, slumping, rolling, cascading, cataracting and centrifuging. Bed motion depends on Froude number, filling degree, wall friction coefficient, ratio of particle to cylinder diameter, angle of internal friction and dynamic angles of repose (Yin et al., 2014). In industrial rotating drums, the rolling or cascading mode is often applied, and the rolling mode is considered optimum for mixing (Demagh et al., 2012; Boateng et al., 2008).

The approach of numerical analysis in multiphase granular flows facilitates understanding of the bed behavior in rotary kilns. CFD simulations can be performed to investigate motions of particles in the bed according to two mathematical models. Both Euler-Lagrange and Euler-Euler models, along with the kinetic theory of granular flow, are used to simulate the bed motion. In the Euler-Lagrange model, the gas phase is treated as a continuum while the solid particles are considered as a discrete phase (Crowe et al., 1998). Trajectories of individual solid particles are calculated to understand the behavior of the kiln bed. The Euler-Euler model considers each phase as a continuum, and continuity and momentum equations for each phase are applied (Valle, 2012).

Both two-dimensional (2D) and three-dimensional (3D) CFD simulations to investigate mixing of one granular phase in a rotary kiln have been reported in the literature. 2D-CFD simulations were done by Liu et al. (2016) to study particle motion and heat transfer in a rotary kiln. The surface particle motion in rotary cylinder was analyzed via a 2D-CFD model to analyze the dynamic characteristics and rheology of a granular viscous flow in a rotary cylinder to validate real cement rotary kiln (Demagh et al., 2012). A three-dimensional study was done by Yin et al. (2014) to understand granular motion during rolling mode in a rotary kiln. The particle residence time and angle of inclination of the rotary kiln were considered in the simulation.

This study focuses on 2D numerical simulations of mixing of two granular phases in a rotating cylinder. Particle motion in a rotating cylinder was considered similar to the bed behavior of the transverse plane in a rotary kiln. The mixing behavior was analyzed considering two granular particle types, of calcium carbonate ($CaCO_3$) and aluminum oxide (Al_2O_3), in rolling mode. A mathematical model based on the Euler-Euler approach and the kinetic theory of granular flow were used to describe the dynamics of particles in the transverse plane. The behavior of $CaCO_3$ and Al_2O_3 was first studied separately, and then mixing of $CaCO_3$ and Al_2O_3 was investigated.

CFD simulations were carried out using ANSYS FLUENT 16.2 and 2D model was developed using ANSYS DesignModeler. The model has the geometrical characteristics of a transverse plane of a rotary kiln.

2 Model Description

2.1 Particle Mixing in a Transverse Plane

The dynamics of the particle bed in a rotating cylinder under rolling mode has been observed by different techniques (Yin et al., 2014). In rolling mode, particles at the top surface of the bed move down continuously while the bottom part moves up showing a plug flow motion. The maximum particle mixing is achieved under rolling mode. Considering the characteristics of the particles movement, the bed motion can be further divided into two regions, an active layer and a passive layer (Boateng et al., 2008). Figure 1 shows a schematic diagram of a kiln operated in rolling mode.

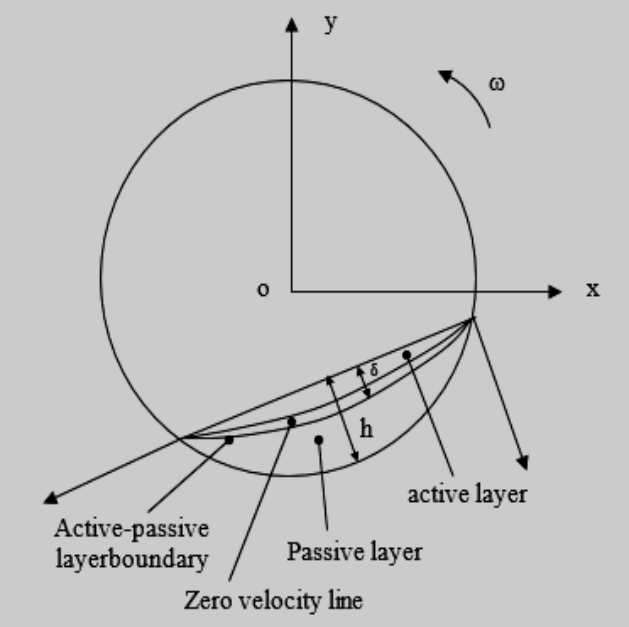


Figure 1. Schematic diagram of a kiln operated in a rolling mode

Here, most of the particle mixing takes place in the active region and mixing in the passive region is negligible. The active layer particle mixing determines the surface renewal rate, which in turn affects the bed-freeboard heat and mass transfer and chemical reactions (Ding et al., 2001). The heat transfer is, however, not included in the present work, which focuses on the particle motion.

2.2 Governing Equation of Two-Fluid Model

2.2.1 Continuity Equations

The continuity equations for the gas phase and the solid phase are as follows:

$$\frac{\partial}{\partial t}\left(\varepsilon_g \rho_g\right)+\nabla\cdot\left(\varepsilon_g \rho_g v_g\right)=0 \tag{1}$$

$$\frac{\partial}{\partial t}\left(\varepsilon_S \rho_S\right)+\nabla\cdot\left(\varepsilon_S \rho_S v_S\right)=0 \tag{2}$$

$$\varepsilon_g+\varepsilon_S=1 \tag{3}$$

Here, ρ is density, v is velocity, ε is volume fraction and t is time. S and g refer to solid phase and gas phase, respectively.

2.2.2 Momentum Equations

Momentum equations describe how the viscous, pressure and gravity forces govern the motion of the gas and the solid particles. The momentum equations for the gas phase and the solid phase are written as:

$$\frac{\partial}{\partial t}\left(\varepsilon_g \rho_g v_g\right)+\nabla\cdot\left(\varepsilon_g \rho_g v_g v_g\right)=-\varepsilon_g \nabla P_g+\varepsilon_g \rho_g g - \beta_{gs}\left(v_g-v_s\right)+\nabla\cdot\left(\varepsilon_g \tau_g\right) \tag{4}$$

$$\frac{\partial}{\partial t}\left(\varepsilon_s \rho_s v_s\right)+\nabla\cdot\left(\varepsilon_s \rho_s v_s v_s\right)=-\varepsilon_s \nabla P_g+\varepsilon_s \rho_s g - \beta_{gs}\left(v_g-v_s\right)+\nabla \tau_s \tag{5}$$

Here P_g , g , β_{gs} and τ_g are the fluid pressure, gravity, drag coefficient between the gaseous and solid phases and viscous stress tensor of the gas phase, respectively.

The viscous stress tensor for the gas phase, τ_g in Eq (4), and for the solid phase, τ_s in Eq (5), are given by the Newtonian form:

$$\tau_g=v_g\left[\nabla v_g+\left(\nabla v_g\right)^T-\frac{2}{3}\mu_g\left(\nabla\cdot v_g\right)\right]I \tag{6}$$

$$\tau_s=\left(-P_s+\zeta_s\nabla\cdot v_s\right)I+\mu_s\left\{\left[\nabla v_s+\left(\nabla v_s\right)^T\right]-\frac{2}{3}\left(\nabla\cdot v_s\right)I\right\} \tag{7}$$

Here P_s , μ_s , ζ_s and I are solid pressure, solid viscosity, solid bulk viscosity and unit tensor, respectively.

Particle-particle collisions create normal forces which are represented by the solid pressure P_s for one solid phase (Benyahia et al., 2000):

$$P_s=\varepsilon_s \rho_s \Theta+2g_0 \rho_s \varepsilon_s^2\left(1+e_p\right)\Theta \tag{8}$$

Here, Θ is the granular temperature (further explained below), g_0 is the radial distribution function and e_p is the particle-particle restitution coefficient.

The bulk viscosity of the solid, ζ_s in Eq (7), is given by (Neri and Gidaspow, 2000):

$$\zeta_s=\frac{4}{3}\varepsilon_s^2 \rho_s d_p g_0\left(1+e_p\right)\sqrt{\frac{\Theta}{\pi}} \tag{9}$$

Here, d_p is the particle diameter.

The solid shear viscosity in Eq (7) is given as (Arastoopour, 2001):

$$\mu_s = \frac{4}{5}\varepsilon_s^2 \rho_s d_p g_0 (1+e_p)\sqrt{\frac{\Theta}{\pi}} + \frac{10\rho_s d_p \sqrt{\pi\Theta}}{96(1+e_p)\varepsilon_s g_0}\left[1+\frac{4}{5}\varepsilon_s g_0 (1+e_p)\right]^2 \quad (10)$$

Wen and Ergun (Huilin and Gidaspow, 2003) proposed that the exchange coefficient β_{gs} between the gas and the solid phase given in Eq (4) and (5) could be calculated by:

$$\beta_{gs}\big|_{Wen\ \&\ Yu} = \frac{3}{4}C_D \frac{\rho_s \varepsilon_s |v_g - v_s|}{d_p}\varepsilon_g^{-2.65} \qquad \varepsilon_g > 0.8 \quad (11)$$

$$\beta_{gs}\big|_{Ergun} = 150\frac{(1-\varepsilon_g)\varepsilon_s \mu_g}{(\varepsilon_g d_p)^2} + 1.75\frac{\rho_g \varepsilon_s |v_g - v_s|}{\varepsilon_s d_p} \qquad \varepsilon_g \le 0.8 \quad (12)$$

The drag coefficient depends on the value of the Reynolds number, Re:

$$\begin{cases} C_D = \dfrac{24}{\mathrm{Re}}\left(1+0.15\,\mathrm{Re}^{0.687}\right) & \mathrm{Re} < 1000 \\ C_D = 0.44 & \mathrm{Re} \ge 1000 \end{cases} \quad (13)$$

$$\mathrm{Re} = \frac{\rho_g \varepsilon_g |v_g - v_s| d_p}{\mu_g} \quad (14)$$

2.2.3 Kinetic Theory of Granular Flow

Granular kinetic theory is extensively used in granular flow modelling to achieve a high level of accuracy of model results to be able to compare with data from the actual system. This theory considers the particle-particle collisions to predict physical properties of the particulate phase. The kinetic theory has been widely used in modelling of fluidized beds to model solid particles in a gas.

A new variable Θ, called the granular temperature, was introduced in this theory. It is a measure of the kinetic energy of the solid. Granular temperature is defined as one-third of the mean square velocity of the random motion of particles $\Theta = v_s'^2/3$, and $v_s'^2$ is the square of the fluctuating velocity of the particle. A transport equation for the granular temperature can be written as (Huilin et al., 2001):

$$\frac{3}{2}\left[\frac{\partial}{\partial t}(\varepsilon_s \rho_s \Theta) + \nabla\cdot(\varepsilon_s \rho_s \Theta) v_s\right] = (\nabla PI + \varepsilon_s \nabla \tau_s) : \nabla v_s + \nabla\cdot(k_s \nabla\Theta) - \gamma_s + \Phi_s + D_{gs} \quad (15)$$

Here, γ_s is dissipation of turbulent kinetic energy, Φ_s is energy exchange between gas and particle and D_{gs} is energy dissipation.

The turbulent kinetic energy dissipation, γ_s in Eq (15), is given as (Neri and Gidaspow, 2000):

$$\gamma_s = 3(1-e_p^2)\varepsilon_s^2 \rho_s g_0 \Theta\left(\frac{4}{d_p}\sqrt{\frac{\Theta}{\pi}} - \nabla\cdot v_s\right) \quad (16)$$

The radial distribution for one solid phase can be expressed as (Rahaman et al., 2003):

$$g_o = \left[1 - \left(\frac{\varepsilon_s}{\varepsilon_{s,\max}}\right)^{1/3}\right]^{-1} \quad (17)$$

The energy exchange between the fluid and solid phases in Eq (15) is defined as (Huilin and Gidaspow, 2003):

$$\Phi_s = -3\beta_{gs}\Theta \quad (18)$$

The rate of energy dissipation per unit volume is in the form of the following equation:

$$D_{gs} = \frac{d_p \rho_s}{4\sqrt{\pi\Theta}}\left(\frac{18\mu_g}{d_p^2 \rho_s}\right)^2 |v_g - v_s|^2 \quad (19)$$

2.3 Simulation

The simulations were performed under rolling mode as this mode is considered to give good mixing. The rotational speed of the cylinder was maintained under a Froude number of 16×10^{-4} to achieve the rolling mode. The cylinder and the particles rotate in the counterclockwise direction.

2.3.1 Physical Properties of Materials and Model Parameters

In this study, two granular phases, made of $CaCO_3$ and Al_2O_3 were used in the simulations. Table 1 shows the related physical properties of gas and solids with model parameters.

Table 1. Physical Properties of Materials and Model Parameters

Parameter	*Description*	*Value*
ρ_{CaCO_3} (kg/m^3)	Particle density	1760
$\rho_{Al_2O_3}$ (kg/m^3)		3000
d_{CaCO_3} (μm)	Particle diameter	175
$d_{Al_2O_3}$ (μm)		1000
f (%)	Degree of particle fill	15
ω (rpm)	Rotational speed	2

2.3.2 Geometry and Mesh

A circle with 0.4m diameter was created in DesignModeler to represent the transverse plane of the rotating cylinder. A mesh was refined to yield about 5500 elements. Figure 2 provides the mesh of the transverse plane.

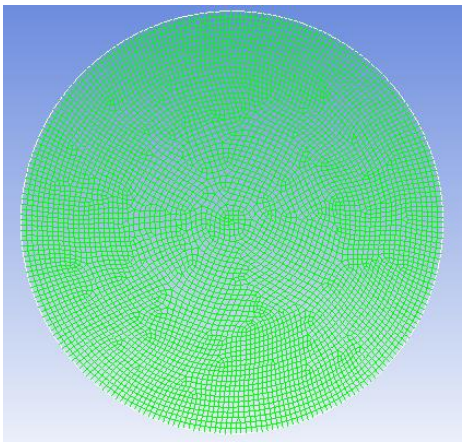


Figure 2. Mesh of the transverse plane

2.3.3 Initial and Boundary Conditions

The main boundary condition of the transverse plane in a rotating cylinder is the relative motion of the bed material and the rotating wall. There, a no-slip condition was assumed, meaning that the relative velocities of the gas and the particles at the wall are set to zero. And it was assumed that particles were subjected to wall friction and gravity.

2.3.4 Solution Strategy and Convergence Criteria

In this study, the finite volume approach was used to solve all the governing equations of the model. Since the flows could be considered incompressible, a pressure-based solver was used. The pressure-velocity coupling was done by a segregated algorithm called "SIMPLE" (Patankar and Spalding, 1972). A second order upwind scheme was used for discretization of the governing equations. The volume fraction was discretized according to the QUICK scheme (Versteeg and Malalasekera, 2007). The time step of the simulations was 10^{-3} s and residual values for the convergence were set to 10^{-3}.

3 Results and Discussion

3.1 Motion of a Single Solid Phase in a Transverse Plane

First, two simulations were performed to understand the bed behavior of $CaCO_3$ and Al_2O_3, respectively. Figures 3(a) and (b) show the volume fraction contours of $CaCO_3$ at 0s and pseudo-steady-state. Initially the top surface is flat, but with the time particles gradually move upwards, following the wall rotation. After a certain time period particles reach a maximum height and then roll down along the top layer in a continuous cyclic motion.

The results indicate that the bed material can be divided into two zones, one active and one passive region. In the active layer, particles move down with relatively high velocities compared to the passive layer.

Figure 4 illustrates the magnitude of the velocity field of $CaCO_3$. Two separated layers can be observed in that a thin upper layer moves with higher velocities than the thick lower layer. Figure 5 shows the velocity contours of $CaCO_3$ particles, which also reveal that the active layer has much higher velocity magnitudes than the passive layer, and the zero velocity region is clearly observed.

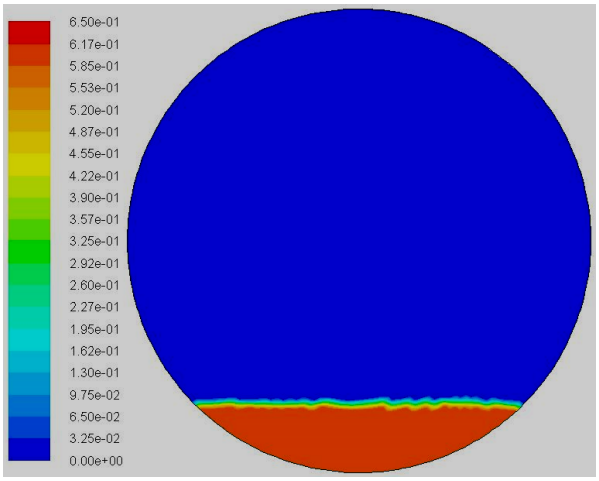


Figure 3(a). Volume fraction contours of $CaCO_3$

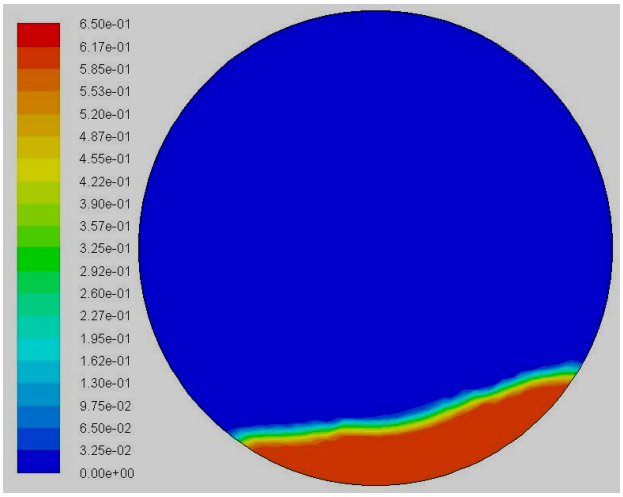


Figure 3(b). Volume fraction contours of $CaCO_3$

In general, the downwards motion in the left direction in the active layer balances the upwards motion in the right direction in the bottom layer, resulting in cyclic pseudo-steady-state flow process. This means that top layer particles exposed to heat transfer from above, as is the case in rotary kilns, would mix with particles in the bottom layer and transfer heat by conduction.

Very similar results were found for the Al_2O_3 motion, so no graphics are included for this particle type.

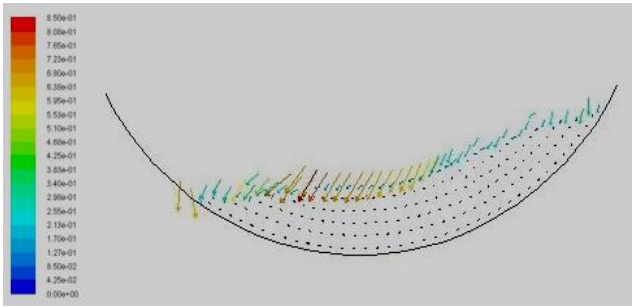


Figure 4. Velocity vector of particle at pseudo-steady-state

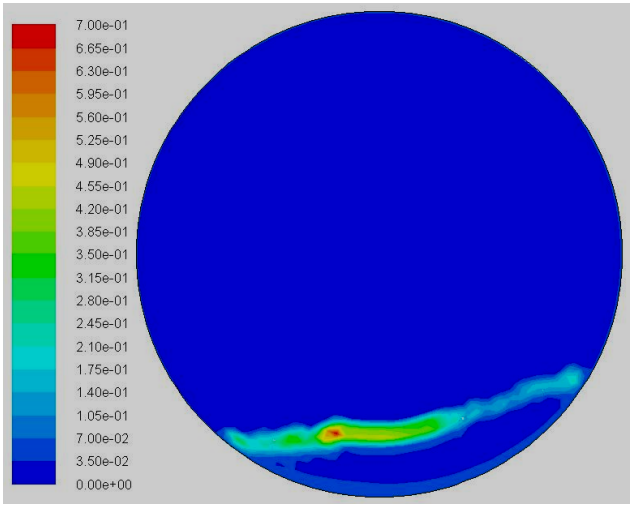


Figure 5. Velocity contours of $CaCO_3$ at pseudo-steady-state

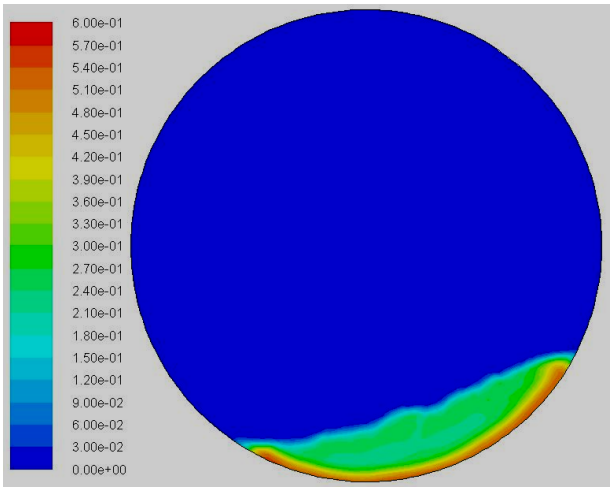


Figure 6. Volume fraction contours of Al_2O_3 in the mixture at pseudo-steady-state

3.2 Mixing of Two Solid Phases in a Transverse Plane

A mixture of $CaCO_3$ and Al_2O_3 were simulated to investigate the mixing performance when two solids with different characteristics are exposed to the rotational motion. Volume fractions of both solids were examined to understand the mixing behavior. Contour plots of volume fractions of solids are shown in Figures 6 and 7. Figure 6 shows the volume fraction variation of Al_2O_3 in the particle bed, revealing that Al_2O_3 accumulates at the bottom of the particle bed. $CaCO_3$, as seen in Figure 7, is collected in the middle of the bed. This illustrates that a mixture of two particle types will undergo segregation instead of mixing during rotation.

Variation of particle size and density are the key factors of segregation in a rotating cylinder. In the active layer of the particle mixture, both solids roll down relative to the slowly moving passive layer. Due to this particle motion, small particles have a higher probability to separate within the active layer and move into middle of the bed. Larger and denser particles move downwards by sustaining the active layer and entering into the passive region at the bottom of the particle bed, near to the wall.

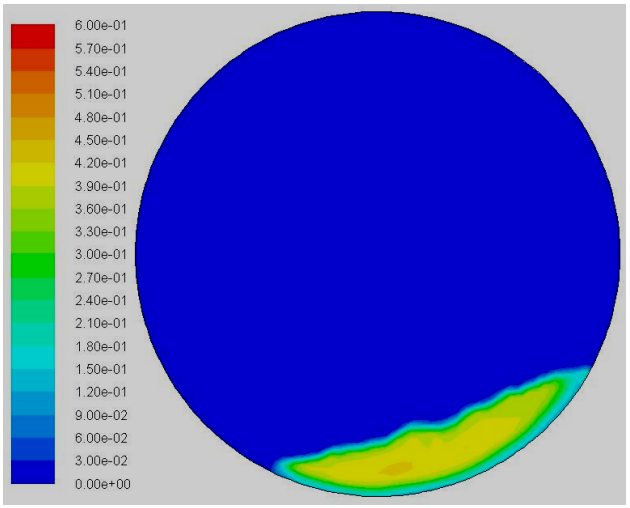


Figure 7. Volume fraction contours of $CaCO_3$ in mixture at pseudo-steady-state

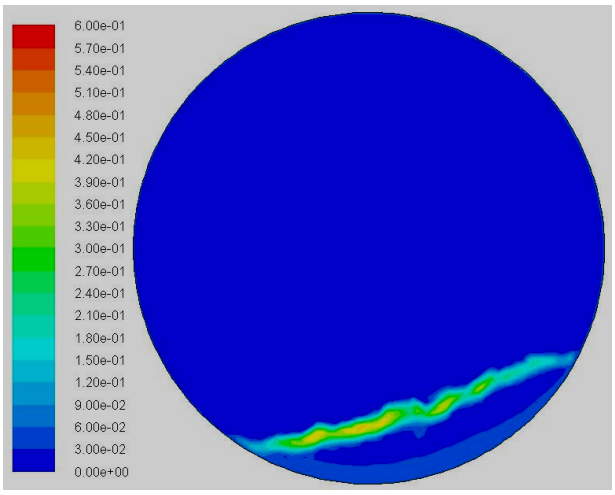


Figure 8. Velocity contours of Al_2O_3 in mixture at pseudo-steady-state

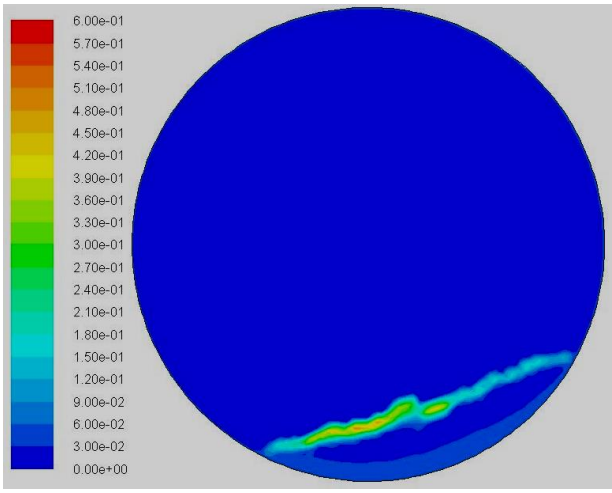


Figure 9. Velocity contours of $CaCO_3$ in mixture at pseudo-steady-state

Velocity contours of both solids are shown in Figure 8 and 9. Both materials are present in the active layer of the mixing bed. The simulation results depict that the intensity of the velocity contours decrease in $CaCO_3$ rather than in Al_2O_3 when materials move towards the lower (left) end of the active layer. This indicates that a smaller amount of $CaCO_3$ particles will remain at the active layer lower end due to segregation.

Generally, particles start to segregate when they are subjected to motion. The particle motion in the active layer facilitates segregation of particles of different size and density.

4 Conclusions

Mathematical modelling of a two-dimensional transverse plane in rotating cylinder, based on the Eulerian approach and the kinetic theory of granular flows, predict the particles mixing behavior in a rotary kiln. The rolling mode can be recommended to achieve internal mixing for a single solid phase. However, particle segregation is observed when two different granular phases are being exposed to the rotary motion under the rolling mode. Lighter and smaller particles are collected in the middle section of the bed while particles with a higher density and size get collected at the bottom of the rotating cylinder. More studies need to be done to understand the mixing mechanism for two granular particles under the rolling mode. In addition to that, further studies are required to determine the best mode for the particles motion in rotating cylinder. Some industrial applications use internal lifters to acquire a higher degree of material mixing. 2-D simulation of the transverse plane with lifters attached to the wall may be applied to investigate to what extent this will improve mixing efficiency of two granular phases.

References

H. Arastoopour. Numerical simulation and experimental analysis of gas/solid flow systems: 1999 Fluor-Daniel Plenary lecture. *Powder Technology*, 119: 59-67, 2001.

S. Benyahia, H. Arastoopour, T. M. Knowlton, and H. Massah. Simulation of particles and gas flow behavior in the riser section of a circulating fluidized bed using the kinetic theory approach for the particulate phase. *Powder Technology*, 112: 24-33, 2000.

A. A. Boateng, *Rotary Kilns: Transport Phenomena and Transport Processes.* USA: Butterworth-Heinemann publications, 2008.

C. Crowe, M. Sommerfeld, and Y. Tsuji, *Multiphase flows with droplets and particles*. USA: CRC Press LLC, 1998.

Y. Demagh, Hocine, B. Moussa, M. Lachi, and L. Bordja. Surface particle motion in rotating cylinders: Validation and similarity for an industrial scale kiln. *Powder Technology*, 224: 260-272, 2012.

Y. L. Ding, J. P. K. Seville, R. Forster, and D. J. Parker. Solid motion in rolling mode rotating drums operated at low to medium rotational speeds. *Chemical Engineering Science*, 56: 1769-1780, 2001.

L. Huilin and D. Gidaspow. Hydrodynamics of binary fluidization in a riser: CFD simulation using two granular temperatures. *Chemical Engineering Science*, 58: 3777-3792, 2003.

L. Huilin, D. Gidaspow, and E. Manger. Kinetic theory of fluidized binary granular mixtures. *Phys. Rev.* E, 64: 061301: 1-8, 2001.

H. Liu, H. Yin, M. Zhang, M. Xie, and X. Xi. Numerical simulation of particle motion and heat transfer in a rotary kiln. *Powder Technology*, 287: 239-247, 2016.

A. Neri and D. Gidaspow. Riser hydrodynamics: Simulation using kinetic theory. *AIChE Journal*, 46: 52-67, 2000.

S. V. Patankar and D. B. Spalding. A calculation procedure for heat, mass and momentum transfer in three-dimensional parabolic flows. *International Journal of Heat and Mass Transfer*, 15: 1787-1806, 1972.

M. F. Rahaman, J. Naser, and P. J. Witt. An unequal granular temperature kinetic theory: description of granular flow with multiple particle classes. *Powder Technology*, 138: 82-92, 2003.

M. A. R. Valle. *Numerical Modelling of Granular Beds in Rotary Kilns*. Master of Science in Computer Simulations for Science and Engineering, Department of Applied Mathematical Analysis, Delft University of Technology, 2012.

H. K. Versteeg and W. Malalasekera, *An introduction to computational fluid dynamics*, second ed. England: Pearson Education Limited 2007.

H. Yin, M. Zhang, and H. Liu. Numerical simulation of threedimensional unsteady granular flows in rotary kiln. *Powder Technology*, 253: 138-145, 2014.

Interactive Visual Analytics of Production Data - Predictive Manufacturing

Juhani Heilala, Paula Järvinen, Pekka Siltanen, Jari Montonen, Markku Hentula, Mikael Haag

VTT Technical Research Centre of Finland Ltd, Espoo, Finland, http://www.vttresearch.com/ {juhani.heilala, paula.jarvinen, pekka.siltanen, jari.montonen, markku.hentula, Mikael.haag}@vtt.fi

Abstract

Manufacturing creates a lot of data, and this is increasing due to digitalization of manufacturing, industrial Internet of Things (IIoT) and needs for product traceability as well as predictive maintenance. Typically data from production material flow is not analyzed and thus the improvement potential is not found. There is need for interactive analytics tools that can turn raw data from heterogeneous data sources e.g. starting from sensor data, manufacturing IT systems, (e.g. Enterprise Resource Planning, ERP, Manufacturing Execution System, MES and Supervisory Control And Data Acquisition, SCADA), into meaningful information and predictions—and presented on easy-to-use interfaces. This paper presents a feasibility study focusing on interactive visual analytics of manufacturing data set carried out at VTT Technical Research Centre of Finland Ltd.

Keywords: manufacturing industry, statistical analysis, machine learning, visual analytics, industrial internet of things

1 Introduction

The role of data in manufacturing has traditionally been understated. Manufacturing generates about a third of all data today (Simafore, 2013), and this is certainly going to increase significantly in the future. Data forms the backbone of all Digital Manufacturing technologies, which will be the centerpiece of the strategy for advancing Manufacturing in the 21st century (Simafore, 2013).

Manufacturing companies are facing global competition – they have to be better, cheaper and faster. In order to manage they need a productivity leap. Manufacturing companies collect huge amount of data of their manufacturing processes. Even though utilization of the data could potentially enable a big productivity leap, this data is poorly used. A recent survey published (MESA, 2016) find out that only 14% of respondents are using manufacturing data in analytics. This is largely because manufacturing companies lack tools and expertise needed to analyze the data. On the other hand, Industrial Internet of Things (IIoT), analytics and simulation methods, collaboration and visualization tools are mature enough to be used and worldwide interest for applying them exists.

Many of current analytic tools require data analytics expertise and are mainly used as a desktop, "island of analysis". Current analytics tools are also static, pre-programmed, focusing mainly on business issues and targeted for upper management level. Typically those tools are expensive and aimed for large organizations. The current tools also have poor synchronous collaborative analysis and decision making features. There are several useful analytics methods that are not included in the current tools.

Advanced analytics refers to the application of statistics, machine learning, data mining and other mathematical methods to manufacturing and business data in order to assess and improve practices.

Predictive analytics is about extracting information from existing data, in order to determine patterns and predicting potential trends and outcomes. Predictive analytics forecasts what might happen in the future. The goal is to go beyond descriptive statistics and reporting on what has happened to providing a best assessment on what will happen in the future. The end result is to streamline decision making and produce new insights that lead to better actions.

In manufacturing, operations managers can use advanced predictive analytics to take a deep dive into historical process data, identify patterns and relationships among discrete process steps and inputs, and then optimize the factors that prove to have the greatest effect on yield.

For networking manufacturers, the IIoT becomes a full ecosystem when software, cloud computing (or in-house servers), and analytics tools are combined to turn raw data into meaningful information or predictions—and when it's presented on easy-to-use interfaces (such as dashboards or mobile Apps) enabling users to monitor, and in some cases, automate response actions or remotely control equipment or systems (PwC, 2015).

The research question is how to convert manufacturing big data to business and manufacturing advantage? In this article we present a feasibility study of applying visual analytics to manufacturing data. For

the purpose we use VTT OpenVA visual analytics platform for measurement data (Järvinen *et al*, 2013). The objectives are twofold: getting experience and guidelines for applying visual analytics in manufacturing and analyzing the feasibility of the VTT OpenVA platform with manufacturing data.

1.1 Impact estimations

Manufacturers taking advantage of advanced analytics can reduce process flaws, saving time and money. Gains will likely show up in both labor productivity and resource productivity: The impact estimations are (MGI, 2011).

- Sensor data-driven operations analytics: -10-20% operation costs, up to +7% revenue
- "Digital Factory" for lean manufacturing: -10-50% assembly cost, + 2% revenue

For extended enterprise real-time visibility between suppliers and the production line allows key value chain participants to optimize material flow and reduce process cycle time. Furthermore, the use of predictive and prescriptive analytics using real-time data allows the enterprise to rectify future bottlenecks and eliminate high costs associated with operational downtime.

General Electric estimates that full-scale exploitation of the industrial internet potential will bring an annual one percentage point increase in global production for the next 15–20 years. One percentage point might not sound much, but calculating one percent growth over fifteen consecutive years as compounding growth we end up with a global increase of ten to fifteen trillion dollars in national product – that is to say, an increase in products and services to the tune of 10,000 or 15,000 billion dollars each and every year. If even a part of this can be realized, we will have gone some way beyond mere hype! (Ailisto 2014).

An efficient workforce with strong data analytical skills and cross-domain expertise can facilitate transition into a smart factory.

2 Visual Analytics in Manufacturing Domain

One key change in the manufacturing process will come in the form of visual analytics (Riley 2015). For many industrial companies, the Industrial Internet of Things (IIoT) is the main source of Big Data. IIoT connects data produced by different objects – such as sensors, devices, machines, humans, other assets, and products – to different applications. IIoT provides access to data generated and manipulated by e.g. intelligent equipment with an IP address, machine-to-machine (M2M) communications, mobility, cloud computing, analytics, and visualization tools.

Inside the factory, having the ability to utilize data masses from orders and machine status allows production managers to optimize operations, factory scheduling, maintenance, and workforce deployment (Noor, 2014).

2.1 Visual Analytics

Visual analytics (Thomas and Cook, 2009; Järvinen *et al*, 2009; Keim *et al*, 2010; Järvinen, 2013,) provides visual and interactive tools to support analytical reasoning and finding insight from data. It combines the human capabilities to interpret visualizations with automatic data processing. A visual analytics tool shows the information in the form of interconnected and interactive visualizations, making the analysis easy for non-experts in data analysis. Behind the visualizations are statistical, data mining and machine learning methods. Users can look for patterns, trends, anomalies, similarities and other relevant features from the visualizations. Visual analytics is an iterative process, (see Figure 1), where users launch analysis, browse and navigate in visualizations, and highlight and select important areas for further study.

The use of visual analytics is still rare in manufacturing operation management. Examples on visual analytics at plant monitoring are shown by (Aehnelt *et al*, 2013; Tack *et al*, 2014).

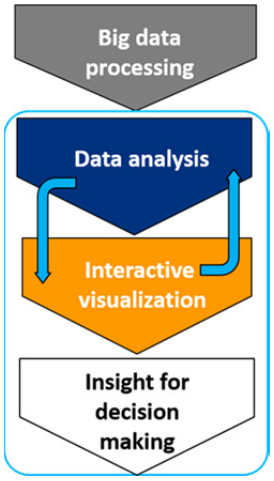


Figure 1. Steps in interactive visual analytics

2.2 Heterogeneous data sources in manufacturing domain

Heterogeneous data stems from various data sources and that comes in a multiplicity of data formats. In the domain of manufacturing, for example enterprise resource planning systems (ERP) are used to manage information about orders and personnel, while manufacturing execution systems (MES) are employed to collect and evaluate data about the production process. For an integrated analysis, different data access interfaces to the different data sources must be used (Aehnelt *et al*, 2013).

In addition to MES and ERP, there are various other systems at factory floor having useful information, such as different sensor systems, SCADA (supervisory control and data acquisition) and various automated systems log files. The most difficult data source is the human at the factory floor, how to get real-time information from him or her. ERP doesn't provide a real time feedback loop from production floor to the planning level. There is latency, delays on submitting the progress data to the ERP and potentially many human interactions are required, thus manual errors are possible. The time stamps from production phases can have inaccuracies. Typically the work phases status are entered at the end of the working shift; there could be the same time for starting and finishing the work phase or some of the work phase recordings are missing.

In case of manual reporting on paper, it requires another human action to type data to the information systems. Information on work progress status, exceptions, missing parts and quality etc. are needed for e.g. real-time control of manufacturing (Järvenpää *et al*, 2014a). Methods for analyzing this historic data are typically missing (Järvenpää *et al*, 2014b).

2.3 VTT OpenVA concept

VTT OpenVA is a visual analytics platform for measurement data by VTT Technical Research Centre Of Finland Ltd. It consists of a data base, a library of visualization and analysis methods, an interactive user interface and a visual analytics engine that delivers data and analysis requests between the different components.

The database stores the application data in a domain independent form. The data base contains data of background variables, measured variables and indicators, and it is populated with application specific metadata.

The analysis and visualization library contains a selection of analysis and visualization methods, and it is extendable. The visual analytics tool adapts itself to each application with the help of the metadata stored into the database. The data to be analyzed is loaded from external sources to the databases through a uniform data interface.

3 Feasibility Test Case – Automated Material Handling System Analysis

The feasibility of OpenVA platform for analyzing manufacturing data was tested with a small sample of industrially relevant data set. The motivation for the feasibility test was to see if the platform would help in the following goal: Advance from descriptive and diagnostic analytics towards predictive analytics – with prescriptive analytics as the next generation capabilities. This could be described as a move from traditional questions of "what happened?" and "why did it happen?" towards a questions "what will happen?" and "how can we make it happen?"

In the feasibility test case, the focus was on visual and interactive analytics, from status monitoring to predictive analytics. The problems of getting data were not studied. The data sets were structured data from a simulation study of an automated material handling system. The feasibility test data was similar to real industrial data that is typically automatically collected from automated equipment and robotics e.g. working time, disturbances, set-up and process times, utilization rate, Overall Equipment Efficiency (OEE) data etc.

In the feasibility test we were utilizing the following parameters from the automated material handling system: capacity (pieces/hour), various equipment (3 machines, 2 robots and lifter) utilization rate, operating time, and storage content (Figure 2). The monitoring of equipment data shows, that production (pieces/hour), utilization rates of robots are low while the utilization rates of the machines are high.

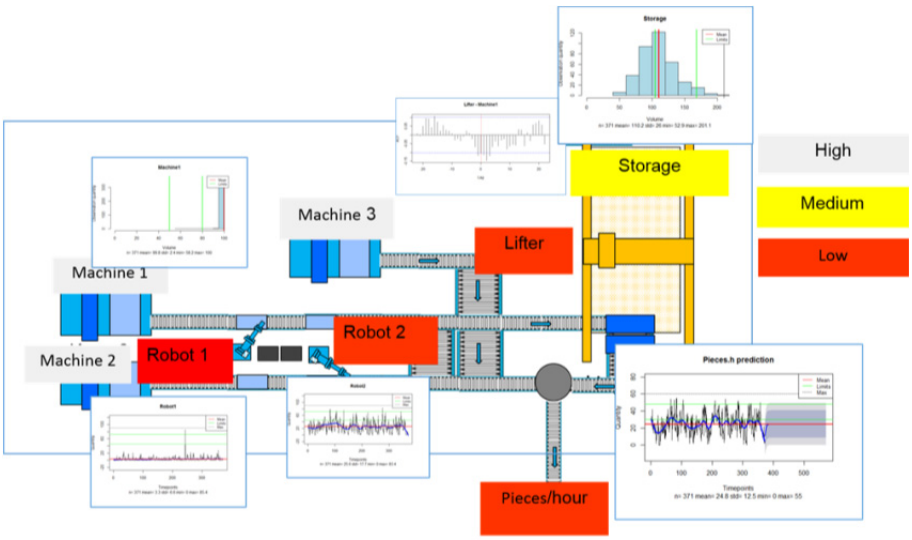


Figure 2. Automated material handling system and equipment's data visualization.

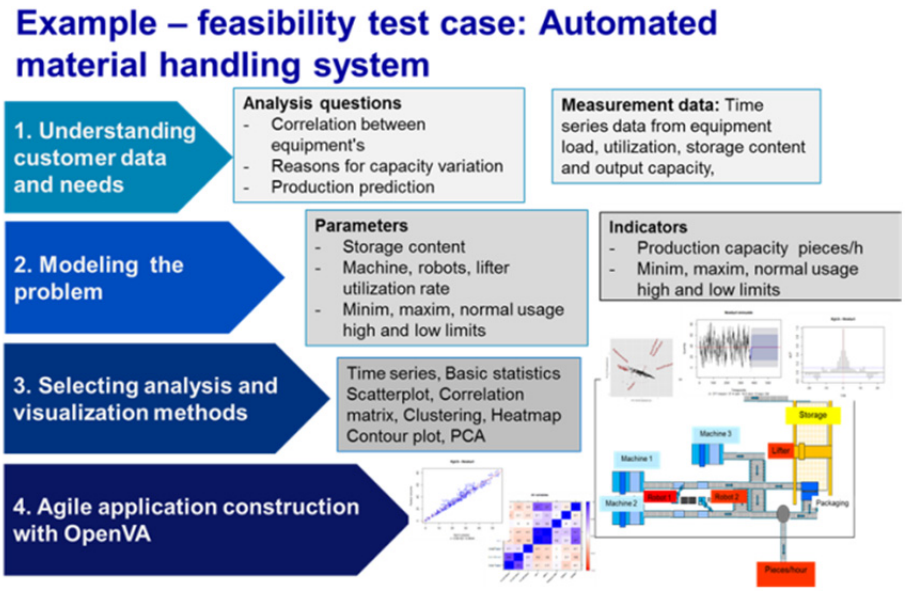


Figure 3. Setting up OpenVA test on automated material handling system

4 Methodology for Setting Up Interactive Visual Analytics

OpenVA is applied by a step-by-step configuration process (**Error! Reference source not found.**). The first step is to understand the customer business, the analysis needs and find out what data is available. Then the next step is to define the phenomena that are followed, to identify the variables that might explain the system behavior and form indicators from the variables. In the third step appropriate analysis and visualization methods are specified. A set of methods is already provided by OpenVA, but new methods can be added. In the final step the analytics application is constructed by configuring the OpenVA platform and loading the application onto the platform database.

In our feasibility study the analysis questions were to study the efficiency of the automated material handling system, to predict the production and find bottlenecks. The selected variables were the utilization rates of each production line component. The indicator chosen for the system production output was finished products/hour. The set of analysis and visualization methods included time series, histogram, contour plot, scatterplot, cross-correlation, correlation matrix, Principal component analysis (PCA), clustering and logistic regression.

The analysis with OpenVA is performed as an iterative reasoning process. First, the user is shown the current status of the performance indicators and the most important variables (Figure 2). The user can study indicators and the other variables in detail with the help of visualizations.

The user starts the analysis by formulating an analysis questions, e.g. "What explains the low production?" Next, the user selects the variables and indicators that might give answers to the questions. The tool suggests suitable analysis and visualization methods to the user based on the number and type of the selected variables. Then the user launches the analysis and gets the results in visual form.

In the analysis of the automated material handling system the most interesting result is shown by the correlation matrix (Figure 4). It shows that the lifter and the production indicator (pieces/hour) have a complete positive correlation, suggesting that the lifter might be the bottleneck of the system. The PCA and logistic regression formula confirm the result. Thereby answers to the analysis questions are: the efficiency of the production line is low, the lifter is the bottleneck of the system and the capacity of the production line is predicted by the lifter alone.

5 Discussion

Predictive analytics applies different analysis methods to predict the possible outcomes of the events that the data describes. It does not give exact measures of what will, could or did happen, just the possibilities of what may occur.

In small series production, different requirements (e.g. high mix, low volume, multiple jobs with different due dates, routing and process time requirements) need to be concurrently processed through the same production setup competing for shared resources of limited capacity. Extreme complexity is characteristic in the discrete made-to-order production, where availability of materials, machines and manpower creates dynamic or moving bottlenecks.

Data quality is a well-known problem in data mining research (Rahm and Do, 2000). The problem is partly caused by missing or erroneous measurements, as well

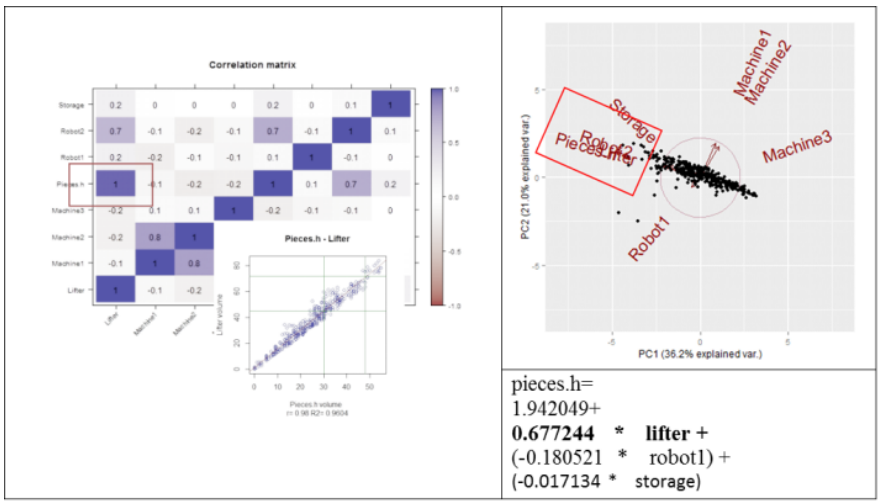


Figure 4. Correlation matrix, cross correlation, PCA and logistic regression

as disparate data formats when combing data from several sources In practical implementations, for example when comparing and estimating measurements from different data collecting applications, the data quality problem is faced immediately. The applications have their own databases, each storing the monitored data in different format. In our test case, the data resulted from a simulation study, and therefore data was cleaner than in a normal industrial case.

The production managers need easy to use tools for finding and eliminating those moving bottlenecks and doing manufacturing process improvements. The production manager needs to consider two different time frames: on the one hand development of equipment and control principles with longer planning and implementation time and on the other hand daily operative decision making. Even if amount of data is low (e.g. because of low production volumes), the predictions do give valuable insight to potential near future events.

As discussed in this paper, the visual and interactive tools can be used to support analytical reasoning and for finding insight from data. However, the visualizations should be chosen case by case so that they are focused on the task at hand, and support exactly those decisions that must be made.

6 Conclusions

In this study our focus was on interactive visual analytics. The key findings regarding the future use of analytics are:

- Instead of looking manually multiple time history plots, numerical tables and reports, one can use the power of interactive visual analytics.
- Getting information from heterogeneous sources, collected from factory floor, manufacturing ICT systems or production engineering design systems for decision making and improvement planning.
- Comparing high and low productivity days, finding correlations, patterns, exceptions, digging deeply to the facts behind events, learning, reasoning and pin pointing improvement areas.
- Visual analytics can diminish the need of multiple customer interviews or simulation studies in a development project by getting findings from existing data.
- Visual analytics can be used with simulation studies, to enhance analysis of the simulation run results in order to get deeper understanding, as done in the feasibility study.

OpenVA platform concept proved out feasible for manufacturing data analysis.

6.1 Future Research

This was a small test with promising results and gives ideas for future research. One of them is synchronous collaborative visual analytics in manufacturing. Analysis and visualization of data are traditionally made asynchronously by individual users with their local tools. Research on synchronous collaborative visual analytics is still in its infancy. Supporting synchronous, multi-party collaboration over networks is becoming increasingly important in order to increase the efficiency of data analysis. An important objective is to combine the best ideas of collaborative work with those of Visual Analytics, i.e. to support interactive collaborative visualizations in multi-party settings. This enables the decision makers to get best possible support from e.g. data scientist without geographical limitations.

For collaborative visualization, integration of data access, communication and messaging functionalities in the Visual Analytics tools are needed. This will support the exchange of opinions and information over

the problem at hand. In addition it requires support solutions for viewing and processing the design collaboratively, linking comments to the specific parts of the object (documents, drawings, 3D designs, visualizations, data sets, etc.), as well as accessing, classifying and filtering those comments at any desired way.

The other development need is reliable and real-time access to data from heterogeneous data sources, from factory floor, sensors, devices, human operators and manufacturing information systems.

For building an IIoT system, all following topics are needed: connectivity, data management, analytics, and interoperability. Reference architectures as well standardization in this domain are evolving e.g. Reference Architecture Model for Industrie 4.0 (RAMI4.0) and the Industrial Internet Reference Architecture (IIRA).

The ability to analyze large amounts of complicated, heterogeneous data with custom-written visual analytics will be key component in the future business and industrial intelligence – analytics. Data-driven decision-making in manufacturing enables productivity leap.

Predictive manufacturing analytics enables users to:

- Progress from monitoring to predictive analytics, optimization and to "how can we make it happen?"
- Near real-time warnings of potential problems, embedded dashboard to factory floor, etc.
- Analyze production characteristics and business performance.

Acknowledgment

Support from VTT Technical Research Centre of Finland Ltd (http://www.vttresearch.com/) is gratefully acknowledged. This small feasibility study was part of VTT spearhead program For Industry. The program was aimed at boosting the competitiveness of the Finnish manufacturing industry.

References

M.Aehnelt, H.Schulz and B.Urban. Towards a Contextualized Visual Analysis of Heterogeneous Manufacturing Data. In G. Bebis et al. (Eds.): *ISVC 2013, Part II, LNCS 8034,* pp. 76–85, Springer-Verlag Berlin Heidelberg, 2013.

H.Ailisto. Industrial Internet - hype or revolution? 2014. Available at https://vttblog.com/2014/05/07/industrial-internet-hype-or-revolution/

E.Järvenpää, H.Tokola, T.Salonen, M.Lanz, M.Koho and R. Tuokko. R. Requirements for Manufacturing Operations Management and Control Systems in a Dynamic Environment. *In Proceedings of the 24th International Conference on Flexible Automation & Intelligent Manufacturing* -FAIM 2014. San Antonio, Texas; University of Texas at San Antonio. Center for Advanced Manufacturing and Lean Systems. 2014a, http://dx.doi.org/10.14809/faim.2014.1135

E.Järvenpää, M.Lanz, H.Tokola, T.Salonen, M.Koho, J.Backman, K.Katajisto and H.Reinilä LeanMES: Tuotannonsuunnittelu ja -ohjaus suomalaisissa valmistavan teollisuuden yrityksissä. Nykytila, haasteet ja tarpeet, 2014. LeanMES project report 2014b

P.Järvinen. A Data Model Based Approach for Visual Analytics of Monitoring Data. Licentiate thesis. Aalto University School of Science. Department of Information and Computer Science. 2013. Available at http://lib.tkk.fi/Lic/2013/urn100763.pdf

P.Järvinen, K.Puolamäki, P.Siltanen and M.Ylikerälä. Visual Analytics. Final report. *VTT WORKING PAPERS* 117. 2009. Available at http://www.vtt.fi/inf/pdf/workingpapers/2009/W117.pdf

P.Järvinen, P.Siltanen, K.Rainio. Framework for Visual Analytics of Measurement Data. INFOCOMP 2013: *The Third International Conference on Advanced Communications and Computation*. ISBN: 978-1-61208-310-0, 2013.

D.Keim, J.Kohlhammer, G.Ellis and F.Mansmann, (eds), Mastering the information age: solving problems with visual analytics. First edition edn. Goslar, Germany: Eurographics association, 2010

MESA International and LNS Research. Survey on Metrics that Matter in the manufacturing world. MANUFACTURING METRICS IN AN IoT WORLD. Measuring the Progress of the Industrial Internet of Things, 2016.

MGI. McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity 2011.

A.K.Noor. Putting Big Data to Work. American Society of Mechanical Engineers (ASME), 2014.

PwC. The Internet of Things: what it means for US manufacturing. 2015. Available at http://www.pwc.com/us/en/industrial-products/next-manufacturing/big-data-driven-manufacturing.html

E.Rahm and H.H.Do. Data cleaning: Problems and current approaches, *IEEE Data Engineering Bulletin 2000,* 23(4): 3-13, 2000.

S.Riley. Visual Analytics – The Future Of Manufacturing Processes. Manufacturing Business Technology. 2015. Available at http://www.mbtmag.com/article/2015/08/visual-analytics-%E2%80%93-future-manufacturing-processes ,

Simafore. How predictive analytics can shape manufacturing of the future. 2013. Available at http://www.simafore.com/blog/bid/118789/How-predictive-analytics-can-shape-manufacturing-of-the-future

T.Tack, A.Maier and O.Niggemann. On Visual Analytics in Plant Monitoring. In J.-L. Ferrier et al. (eds.), *Informatics in Control, Automation and Robotics, Lecture Notes in Electrical Engineering 283*, 2014. Doi: 10.1007/978-3-319-03500-0_2,

J.Thomas and K.Cook. Illuminating the path: The research and development agenda for visual analytics. 1st ed. Los Alamitos, 2009.

Cost Optimization of Absorption Capture Process

Cemil Sahin Lars Erik Øi

Department of and Process, Energy and Environmental Technology,
University College of Southeast Norway, Norway, lars.oi@usn.no

Abstract

In this work, a CO_2 absorption process using aqueous monoethanol amine (MEA) as solvent for a post combustion capture plant was simulated using Aspen HYSYS. An Aspen HYSYS spreadsheet was used for equipment dimensioning, cost estimation and cost optimization. A standard process and a vapor recompression process for 85 % CO_2 removal were simulated using the Li-Mather thermodynamic model. The energy consumptions and the total cost were calculated and compared. Cost optimum process parameters were calculated from sensitivity analysis. The vapor recompression process was shown to be both energy and cost optimum. With 20 years calculation period, the cost optimum absorber packing height was 16 meter, optimum temperature approach was 14 K and optimum recompression pressure was 130 kPa. With 10 years calculation period, the optimum values for the same parameters were 16 meter, 17 K and 140 kPa. Calculations of optimum process parameters dependent on factors like the calculation period have not been found in literature. Except from the temperature approach, the optimum values varied only slightly when the calculation period was changed.

Keywords: CO_2, amine, absorption, cost estimation, Aspen HYSYS

1 Introduction

CO_2 absorption with aqueous monoethanol amine (MEA) as solvent is well-known and mature for large scale post combustion CO_2 capture. This process has a high energy consumption. There have been suggested many modifications to reduce the heat demand for the reboiler in the desorption column. Vapor recompression is one of the simplest way to reduce the energy need (Øi et al., 2014). In this paper, the standard and vapor recompression configuration were simulated, cost estimated and optimized using the process simulation program Aspen HYSYS. The results in this work are mainly from a Master Thesis work (Sahin, 2016).

The economic performance of a CO_2 removal process plant has been evaluated in earlier work (Abu-Zahra et al., 2007). The amine concentration, lean amine loading and desorber column pressure was determined to be the main factors influencing on the cost. The capital cost and energy consumption of different configurations have been evaluated using the Unisim and ProTreat simulation programs (Karimi et al., 2011). One reference (Fernandez et al., 2012) has performed cost estimation and found optimum cost parameters using rate based Aspen Plus simulation. Another reference (Cousins et al., 2011) simulated different configurations using Aspen Plus and compared the energy consumption.

There is very limited published work on cost optimum parameters, cost evaluation and optimization of alternative configurations. At University College of Southeast Norway, different modification alternatives were simulated and cost estimated using Aspen HYSYS (Øi et al., 2014). The cost optimum absorber packing height, the minimum temperature approach in the main heat exchanger and the gas inlet temperature were calculated for the standard process using Aspen HYSYS (Øi, 2012; Kallevik, 2010).

The aim of this work is to simulate and cost estimate the standard CO_2 absorption process and the vapor recompression process. Then different parameters are varied to energy optimize and cost optimize different process parameters and the total process. A special aim in this work is to vary cost factors, and especially the calculation period between 10 and 20 years to evaluate the effect on the optimum parameters.

2 Process description

2.1 Principles of Standard Process

The standard amine based CO_2 capture process contains an absorber, a stripper with reboiler and condenser, a lean/rich heat exchanger, pumps and a lean amine cooler as shown in Figure 1.

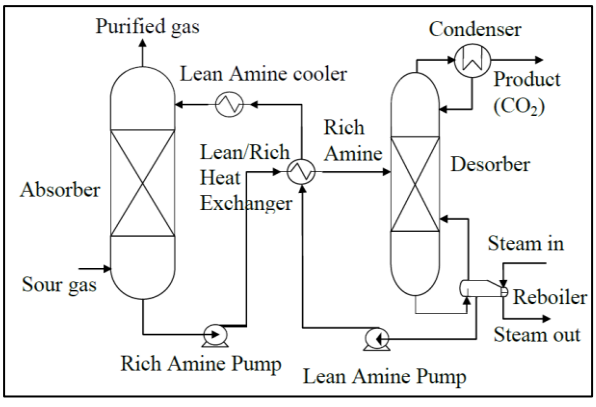


Figure 1. Principle for standard amine based CO_2 capture process (Aromada and Øi, 2015).

When the flue gas rises from the bottom to the top of the absorption column, CO_2 is absorbed by the solvent. The rich amine is pumped from the absorption column to the desorber column passing through the lean/rich heat exchanger. In the stripper, the absorbed CO_2 is removed from the solvent using thermal energy supplied to the reboiler. The lean amine from the bottom of desorber is pumped to the absorption column via the lean/rich heat exchanger and the lean amine cooler.

2.1 Principles of Vapor Recompression

The principles of vapor recompression are shown in Figure 2. The difference from the standard process is that the regenerated amine solution from the desorber is flashed using a valve and led to a two-phase separator. The liquid from the separator is returned back to the absorber. The vapor from the separator is compressed and sent back to the bottom of the desorber.

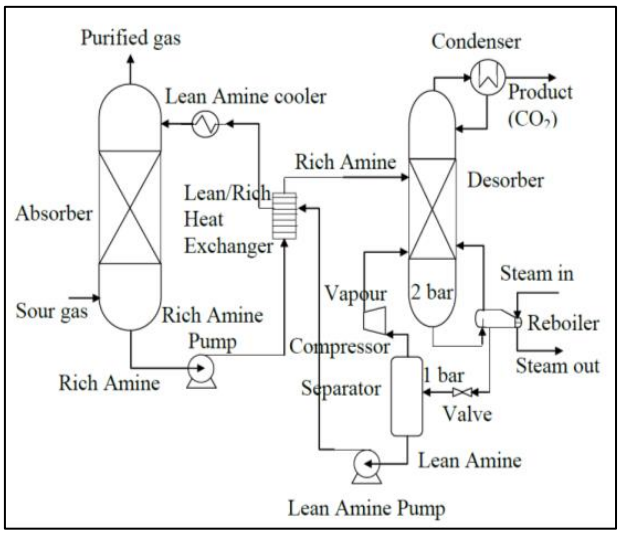


Figure 2. Principle for vapor recompression process (Aromada and Øi, 2015).

3 Models

3.1 Equilibrium Models

The Kent-Eisenberg (Kent and Eisenberg, 1976) and the Li-Mather (Li and Mather, 1996) vapor/liquid equilibrium models are available models in the Amine Property Package in Aspen HYSYS. Both models are quite complex involving several adjusted parameters. The Kent-Eisenberg model is claimed to give faster convergence while the Li-Mather model is claimed to be more robust (Øi et al., 2014). The non-ideal vapor phase model is used in the simulations in this work.

3.2 Column Models and Iteration Algorithms

Equilibrium stages are used to model the columns. A certain packing height can be modelled as an equilibrium stage. One equilibrium stage can be calculated with the assumption that there is equilibrium between the CO_2 concentration in the gas and liquid leaving the stage. A Murphree efficiency can be used to model deviation from equilibrium. It can be specified explicitly for each stage in a column in Aspen HYSYS (Øi, 2007). The Modified HYSIM Inside-Out algorithm with adaptive damping is specified for the columns. This is also done in earlier simulations (Øi, 2007).

4 Process Simulations

4.1 Specifications for Standard CO_2 Capture Process

The standard CO_2 removal process has been simulated in Aspen HYSYS with the specifications in Table 1. The Aspen HYSYS process flow diagram is shown Figure 3. The specifications are based on earlier works with amine absorption from a natural gas based power plant by Øi (2007). The amine package with the Li-Mather model was used in all the simulations in this work.

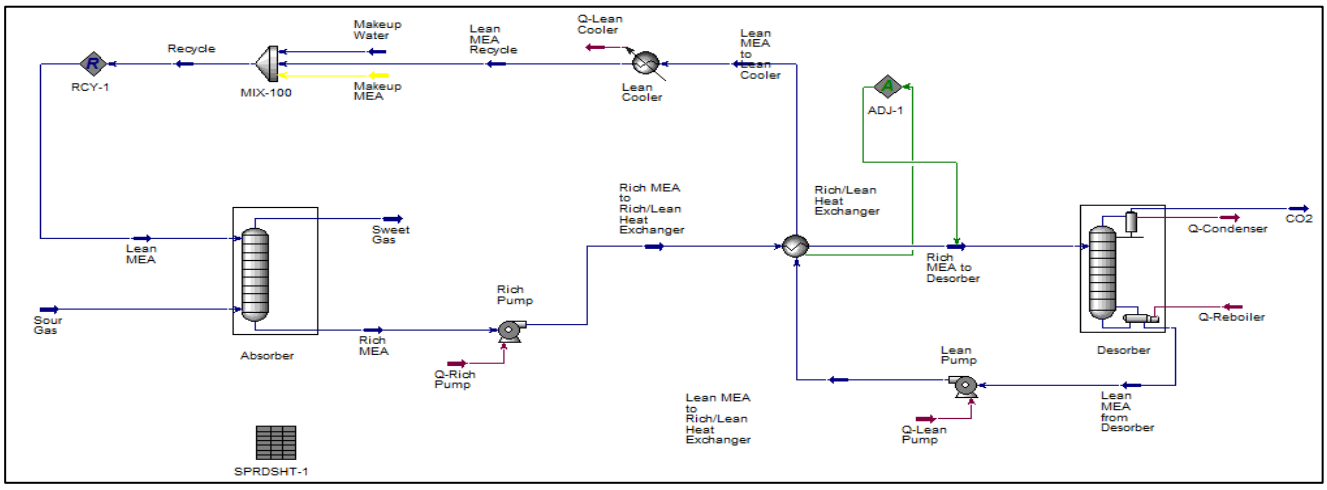


Figure 3. Aspen HYSYS flow-sheet for standard CO_2 capture process.

Table 1. Specifications for standard process.

Parameter	Unit	Value
CO_2 removal grade	mass %	85
Inlet gas temperature	℃	40
Inlet gas pressure	kPa	110
Inlet gas flow rate	kmol/h	85000
CO_2 in inlet gas	mole %	3.73
Water in inlet gas	mole %	6.71
Nitrogen in inlet gas	mole %	89.56
Lean amine temperature	℃	40
Lean amine pressure	kPa	101
Lean amine rate	kmol/h	128000
MEA content in lean amine	mass %	29
CO_2 content in lean amine	mass %	5.5
Number of stages in absorber	-	16
Murphree efficiency in absorber	-	0.15
Rich amine pump pressure	kPa	200
Rich amine temperature to desorber	℃	104.5
Number of stages in desorber	-	8 (2+6)
Murphree efficiency in desorber	-	1
Reflux ratio in stripper	-	0.3
Reboiler temperature	℃	120
Minimum ΔT in Rich/Lean HX	℃	10

The heat consumption in the reboiler was calculated to 3.72 MJ/kg CO_2 which is slightly higher than in some references (Karimi et al., 2011; Fernandez et al., 2012; Øi, 2007) but lower than in some other references (Cousins et al., 2011; Øi and Vozniuk, 2010). The range in these references are from 3.56 to 3.80 MJ/kg.

4.2 Simulation of Vapor Recompression

A simulation of the vapor recompression process was performed. The Aspen HYSYS flow diagram is in Figure 4. The flash pressure was specified to 120 kPa and the efficiency of the compressor was defined to 75 %. To achieve 85 % CO_2 removal efficiency, the lean amine flow rate was adjusted to 111000 kmol/h and the resulting lean amine CO_2 concentration was 5.12 mass %. The rich amine temperature to the desorber was adjusted to 95.8℃ to achieve the minimum temperature approach as 10℃ in the heat exchangers.

With the vapor recompression modification the reboiler heat consumption was reduced from 3.72 MJ/kg to 3.02 MJ/kg. In literature the calculated reboiler heat consumption was 3.03 and 3.04 MJ/kg (Fernandez et al., 2012; Cousins et al., 2011).

The equivalent heat consumption was calculated to 3.28 MJ/kg using a conversion efficiency from reboiler heat (low pressure steam) to electricity as 25 %. The equivalent heat consumption was calculated in literature to 3.30 MJ/kg (Fernandez et al., 2012).

4.3 Dimensioning and Cost Estimation

The absorption and desorption column diameters were calculated based on gas velocities of 2 m/s and 1 m/s respectively. The packing height was determined with the assumption of 1 meter structured packing for each stage. The column height in addition to packing was 24 m for the absorber and 20 m for the desorber. The pressure drop for each stage in the absorber was 900 Pa.

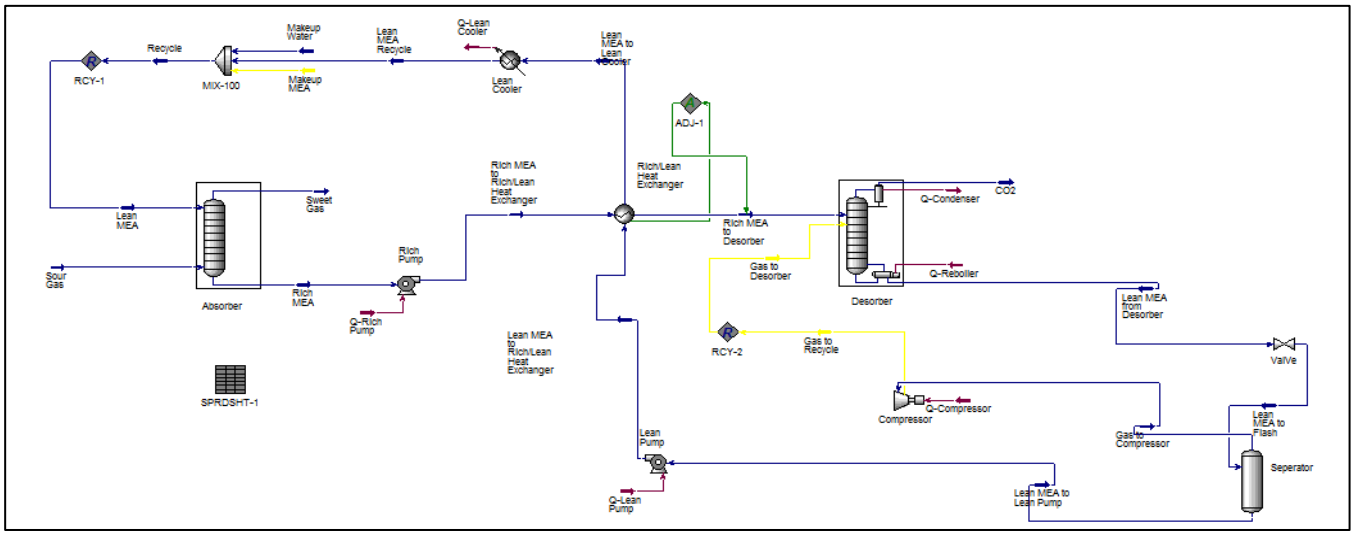


Figure 4. Aspen HYSYS flow-sheet for vapor recompression process.

The heat exchangers were specified as plate heat exchanger. The overall heat transfer coefficients were calculated for the lean/rich heat exchanger and lean amine cooler and estimated to 2500 W/m^{2}K for the reboiler and the condenser. Pumps and compressors were specified with 75 % adiabatic efficiency.

The operating cost was mainly estimated from the energy cost. The maintenance cost was specified as 5 % of the capital cost. The rate of currency was 1 USD to 8.6 NOK. The electricity cost and the steam cost was specified to 0.62 and 0.155 NOK/kWh based on a conversion efficiency from low pressure steam to electricity of 25 %. Operating time per year was defined to 8000 hours. The calculation period was 20 years and the rate of interest was 7 %.

Open source internet cost estimation calculators, one is based on data from Peters and Timmerhaus, were used to calculate the equipment cost (Milligan and Milligan, 2014; Peters et al., 2002). Outside the range of equipment size, power law exponents of 0.57, 1.0, 0.55, 0.28 and 0.95 were used for column vessels, packing, heat exchangers, pumps and compressors. The Chemical Engineering Plant Cost Index was used to convert to USD (2015). After finding the equipment cost, the installed equipment cost was calculated using the detailed factor method using factors for engineering cost, administration cost and contingency.

The net present value using the operating and total installed cost was calculated to 3810 MNOK for the standard process and 3540 MNOK for the vapor recompression process.

5 Cost Optimization

After concluding that the vapor recompression process was more cost optimum than the standard process, the net present value was calculated with varying conditions to find the optimum parameters for the vapor recompression process.

5.1 Cost Optimization of Minimum Temperature Approach

The trade-off was performed with varying minimum approach temperature in the lean/rich heat exchanger. The heat exchanger area and the steam consumption in the reboiler were the main affected variables. The energy consumption and cost optimum as a function of minimum approach temperature is shown in Figure 5.

The calculated optimum value was 14°C with the net present value of 3490 MNOK. As the minimum approach temperature increased, the operational cost increased and the investment cost decreased continuously. Similar calculations were performed for a reduced calculation period of 10 years. The cost optimum temperature approach is shown in Figure 6. The optimum value was 17°C with a net present value of 2710 MNOK. The reason for the higher value is that the investment cost dominates more than the operational cost as the calculation period decrease.

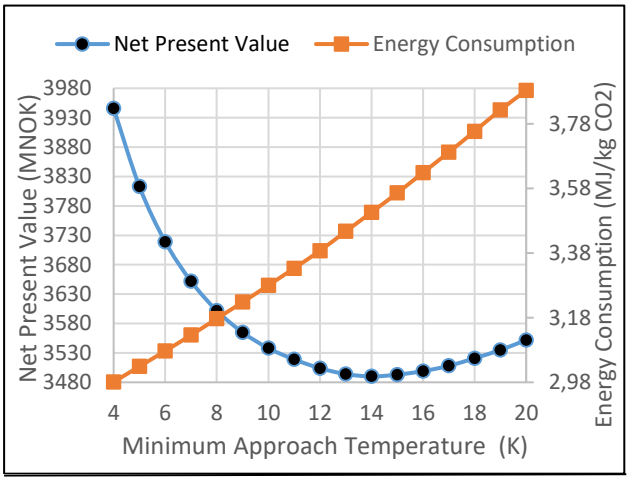


Figure 5. Net present value and energy consumption as a function of minimum temperature approach in heat exchanger for 20 years calculation period.

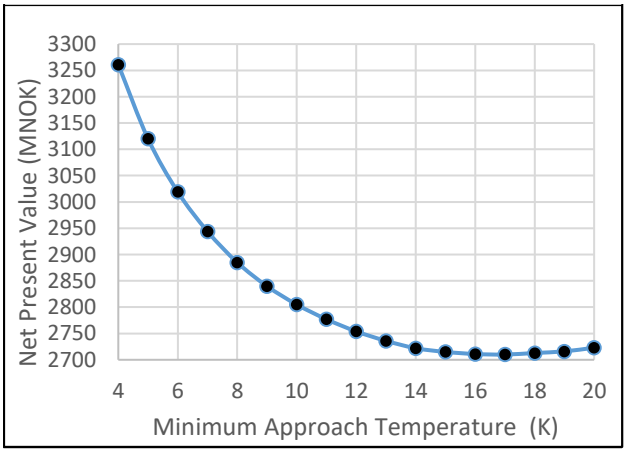


Figure 6. Net present value and energy consumption as a function of minimum temperature approach in heat exchanger for 10 years calculation period.

The cost optimum temperature approach is comparable to values found in literature. Comparisons of optimum temperature approach for different calculation periods have however not been found in literature. The optimum was calculated to 12°C for the calculation period of 15 years and the discount rate of 10.5 % (Øi et al., 2014) and to 19°C for the calculation period of 10 years and an interest rate of 7 % (Øi, 2012).

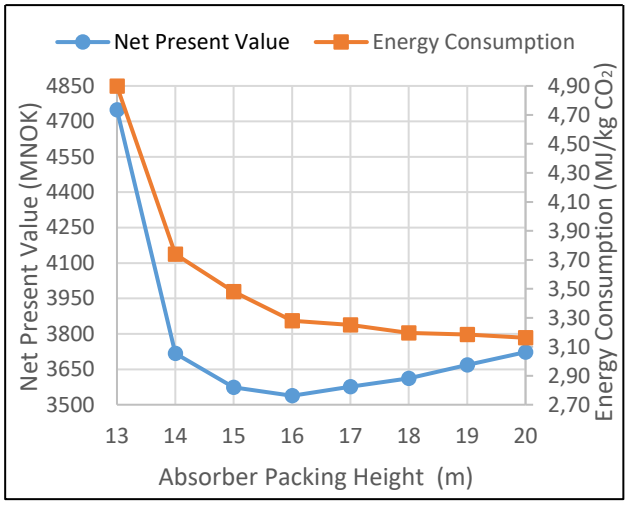


Figure 7. Net present value and energy consumption as a function of absorber packing height for 20 years.

5.2 Cost Optimisation of Absorber Packing Height

The number of absorber stages equivalent to 1 meter was varied between 13 and 20. The cost optimum value of 3540 MNOK was achieved with 16 stages as shown in Figure 7. As the number of stages increased, the necessary amine flow rate to keep the 85 % CO_2 removal efficiency decreased. The equivalent heat consumption decreased from 4.9 MJ/kg CO_2 at 13 stages to 3.16 MJ/kg CO_2 at 20 stages. The investment of the absorber column, the amine heat exchanger, the compressor, the electric consumption due to fan and compressor and steam consumption in the reboiler were the major changes. For 10 years calculation period, the cost optimum value was also achieved with 16 stages as shown in Figure 8.

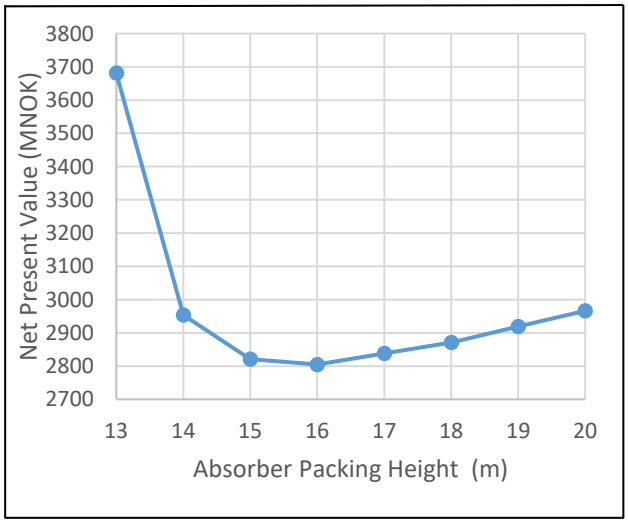


Figure 8. Net present value and energy consumption as a function of absorber packing height for 20 years.

Similar values are found in literature. For the calculation period of 15 years, 16 stages in the absorption column was calculated as the cost optimum value (Øi et al, 2014). In another optimization calculation, the optimum value was also 16 stages for a calculation period of 10 years (Øi, 2012). Evaluations of the influence of the calculation period on the optimum packing height have not been found in literature.

5.3 Cost Optimization of Flash Pressure

The flash pressure (pressure before recompression) can be varied. The cost of the lean/rich heat exchanger, the compressor, the reboiler and the energy consumption in the compressor and the reboiler changed when the flash pressure was varied. The equivalent heat consumption was reduced to 3.27 MJ/kg CO_2 at 1.1 bar. The cost optimum pressure was calculated to be 1.3 bar with a net present value of 3530 MNOK. The result of the cost optimum flash pressure is given in Figure 9. The cost optimum flash pressure is slightly different from the energy optimum value. The reason is that the energy saving is dominated by the investment cost of the compressor until 1.3 bar.

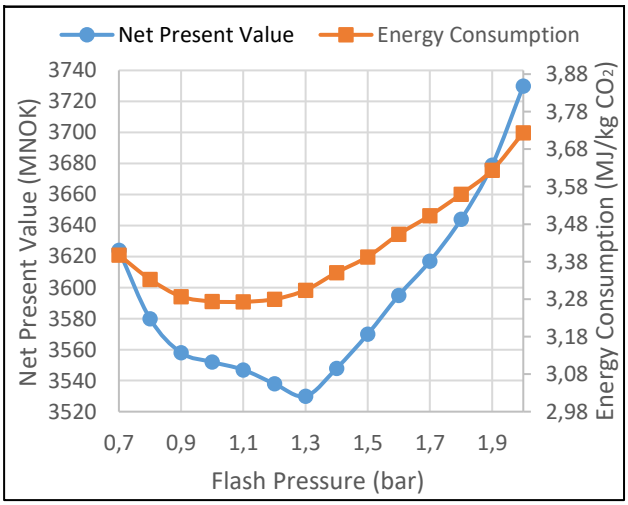


Figure 9. Net present value and energy consumption as a function of flash pressure for 20 years.

For a calculation period of 10 years, the cost optimum flash pressure was calculated to be 1.4 bar as shown in Figure 10. The dependency of the optimum flash pressure on the calculation period is not obvious. It is however reasonable that as the calculation period increases, the optimum value will intend to decrease because the operational cost becomes more dominant.

In literature, energy optimum flash pressures have been calculated to 1.12 and 1.17 bar (Karimi et al., 2011). In Fernandez et al. (2012), the cost optimum flash pressure was calculated to 1.2 bar. The cost optimum

pressure is slightly higher than the energy optimum because the change in compressor cost is significant. Calculations of optimum flash pressure dependent on factors like the calculation period have not been found in literature.

5.4 Evaluation of Uncertainty

The calculations of the equilibriums, the material balances and the energy consumptions in the process simulations are regarded to be reasonable accurate. As a result of this, the deviation in calculated energy consumptions compared to values found in literature is quite low.

The uncertainties in cost estimation of the equipment are much larger. First, there are uncertainties in the dimensioning of the process equipment. Then there is a high uncertainty in the cost of especially heat exchangers and the absorption columns. And the installation cost of all types of equipment also have high uncertainty. The chosen calculation period and chosen discount rate will also influence on the total cost estimate.

The main aim in this paper was to find cost optimum process parameters. It is of interest to find out whether these optimums are dependent on the choice of different cost factors. When comparing optimum parameters calculated in this work compared to values found in literature, the deviation is rather low.

In this paper, optimum parameters have been calculated for a calculation period of 10 and 20 years. The differences in calculated optimums are very small. The only deviation was in the optimum temperature difference approach in the main heat exchanger that changed slightly from 14 to 17 °C.

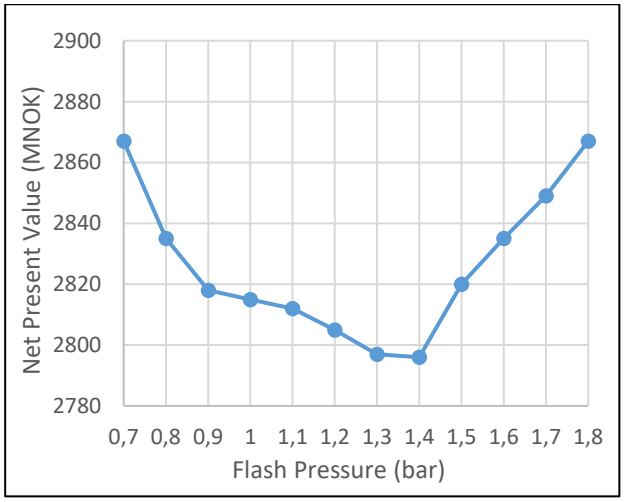


Figure 10. Net present value as a function of flash pressure for 10 years.

6 Conclusions

Simulations and optimizations of an amine-based CO_2 removal process were performed in the search for a cost optimum process. The process configurations examined were a standard process and a vapor recompression process. The energy consumptions and the total cost were calculated and compared.

The vapor recompression modification gave lower total cost compared to a standard process. Optimizations of parameters like minimum temperature approach, height of absorber packing and flash pressure were performed. Cost optimum process parameters were calculated from sensitivity analysis.

Calculations of optimum process parameters dependent on factors like the calculation period have not been found in literature. With 20 years calculation period, the cost optimum absorber height was 16 meter, optimum temperature approach was 14 K and optimum recompression pressure was 130 kPa. With 10 years calculation period, the optimum values for the same parameters were 16 meter, 17 K and 140 kPa. Except for the minimum temperature approach, it seems like when varying different cost factors like the calculation period, the cost optimum values vary only slightly.

References

M. R. M. Abu-Zahra, J. P. M. Niederer, P.H.M. Feron and G. F. Versteeg. CO_2 capture from power plants: Part II. A parametric study of the economical performance based on mono-ethanolamine. *International Journal of Greenhouse Gas Control*, 1:135-142, 2007.

S. A. Aromada and L. E. Øi. Simulation of Improved Absorption Configurations for CO_2 Capture. *In Linköping Electronic Conference Proceedings from SIMS 56*, pages 21-29, 2015.

A. Cousins, L. T. Wardhaugh and P. H. M. Feron. Preliminary analysis of process flow sheet modifications for energy efficient CO_2 capture from flue gases using chemical absorption. *Chemical Engineering Research and Design*, 89:1237-1251, 2011.

E. S. Fernandez, E. J. Bergsma, F. de Miguel Mercader, E. L.V. Goetheer and T. J. H. Vlugt. Optimisation of lean vapour compression LVC as an OPTION for post-combustion CO_2 capture: Net present value maximisation. *International Journal of Greenhouse Gas Control*, 11:114-121, 2012. doi: http://dx.doi.org/10.1016/j.ijggc.2012.09.007

O. B. Kallevik. *Cost estimation of CO_2 removal in HYSYS*. Master Thesis, Telemark University College, Porsgrunn, Norway, 2010.

M. Karimi, M. Hillestad and H. F. Svendsen. Capital costs and energy considerations of different alternative stripper configurations for post combustion CO_2 capture. *Chemical Engineering Research and Design*, 89:1229-1236, 2011.

R. L. Kent and B. Eisenberg. Better data for amine treating. *Hydrocarbon processing*, 55(2):87-90, 1976.

Y. Li and A. E. Mather. Correlation and prediction of the solubility of CO_2 and H_2S in aqueous solutions of triethanolamine. *Industrial & Engineering Chemistry Research*, 35:4804-4809, 1996.

D. A. Milligan and J. A. Milligan. *Internet cost estimation program*. 2014. Available (06.02.2016): http://www.matche.com/equipcost/Default.html

M. S. Peters, K.D. Timmerhaus and R. E. West. *Internet cost estimation program*. 2002. Available (01.02.2016): http://www.mhhe.com/engcs/chemical/peters/data/

C. Sahin. *Optimization of CO_2 capture based on cost estimation.* Master Thesis, University College of Southeast Norway, 2016.

L. E. Øi. Aspen HYSYS simulation of CO_2 removal by amine absorption from a gas based power plant. *In The 48th Scandinavian Conference on Simulation and Modelling (SIMS2007)*, Göteborg, Sweden, 2007.

L. E. Øi. *Removal of CO_2 from exhaust gas.* PhD Thesis, Telemark University College, Porsgrunn. (TUC 3: 2012)

L. E. Øi, T. Bråthen, C. Berg, S. K. Brekne, M. Flatin, R. Johnsen, I. G. Moen and E. Thomassen. Optimization of configurations for amine based CO_2 absorption using Aspen HYSYS. *Energy Procedia*, 51:224-233, 2014.

L. Øi and I. Vozniuk. Optimizing CO_2 absorption using split-stream configuration. *In conference proceedings of Processes and Technologies for a Sustainable Energy*, Ischia, Italy, 2010.

Fuzzy Modelling of Air Preparation Stage in an Industrial Exhaust Air Treatment Process

Aleš Šink[1] Gašper Mušič[2]

[1]Inea d.o.o., Slovenia, ales.sink@inea.si
[2]Faculty of Electrical Engineering, University of Ljubljana, Slovenia, gasper.music@fe.uni-lj.si

Abstract

The paper is focused on practical aspects of advanced nonlinear identification method applied to a real industrial process. Fuzzy identification is used to model the air preparation stage within a system for reducing nitrogen oxides (NOx) emissions in exhaust air from the dryers and ovens in a factory of automotive catalytic converters. The system for NOx emissions reduction operates efficiently in predetermined temperature and air flow ranges of the exhaust air only. Due to those conditions, exhaust air from the dryers and ovens must be prepared in advance by controlling the ventilator speed and fresh air and exhaust air dampers positions. At the same time operating conditions of dryers and ovens have to be maintained within defined ranges. Currently used control system of the exhaust air preparation shows some deficiencies, so a feasibility study of possible improvements has been carried out. Modelling presented in this paper has been used to evaluate and compare control solutions. The results show such an improvement is feasible. The proposed control system can be ready for implementation in the real process with minor changes of the controller parameters and supervisory logic settings.

Keywords: fuzzy logic, Takagi-Sugeno model, catalytic converter, emission reduction, process control

1 Introduction

Emissions of pollutants are a challenging problem in many contemporary industrial processes. In particular, emissions of atmospheric pollutants in exhaust air have direct influence on quality of the living conditions in the neigbouring areas of industrial plants, as well as other important environmental impacts, e.g. climate change.

The nitrogen oxides (NOx) emissions in the exhaust air can be effectively reduced by catalytic converters, which are based on the same operation principles as used in internal combustion engine exhaust system in traffic vehicles. Industrial catalytic converters typically work on selective catalytic reduction (SCR) principle, although also selective non-catalytic reduction (SNCR) based systems have been used, e.g. in waste incineration plants. With SCR, a gaseous reductant is added to the stream of exhaust air, typically anhydrous ammonia, aqueous ammonia or urea, which reacts with gas mixture in the exhaust air and the catalyst to form molecular nitrogen, water and carbon dioxide.

To achieve the (near) optimal operation of the catalytic converter unit, the incoming production process exhaust air must be properly conditioned. This involves control of flow, pressure and the temperature.

The currently used control system of the exhaust air preparation stage in the process under investigation shows some deficiencies, so a feasibility study of possible improvements has been carried out. A substantial part of the feasibility study was development of a mathematical model, which is presented in this paper. The model was necessary in order to be able to experiment with control system. The presentation focuses on practical aspects of deriving a model by fuzzy identification based on actual production data.

2 Air preparation process

The system considered in the paper is a part of the catalytic converters production plant. The produced catalytic converters are used to reduce NOx emissions from internal combustion engine driven vehicles.

Due to the used production technology, the NOx emissions are present in the production process itself so an emissions reduction unit is installed. The operating conditions are varying substantially with the changing production. Therefore an air preparation stage is installed in between the production process and the catalytic converter (Fig. 1). This way a stable operating regime of the NOx emissions reduction unit is achieved.

The process contains a pipe system delivering exhaust air from ovens and dryers, equipped with controllable dampers and ventilators; a set of pressure/differential pressure and temperature sensors, and damper position sensors is installed. Low pressure is maintained to transport the air out of the process and the air flow is calculated from differential pressure reading within the catalyst converter unit, taking into account air pressure and air temperature. According to specifications the control relevant signals are: air flow, pressure before ventilator, differential pressure over the ventilator, ovens exhaust pressure and dryers exhaust pressure.

The currently operating system involves automatic control of the exhaust air preparation system, but the control system is not operating in a closed-loop manner. The con-

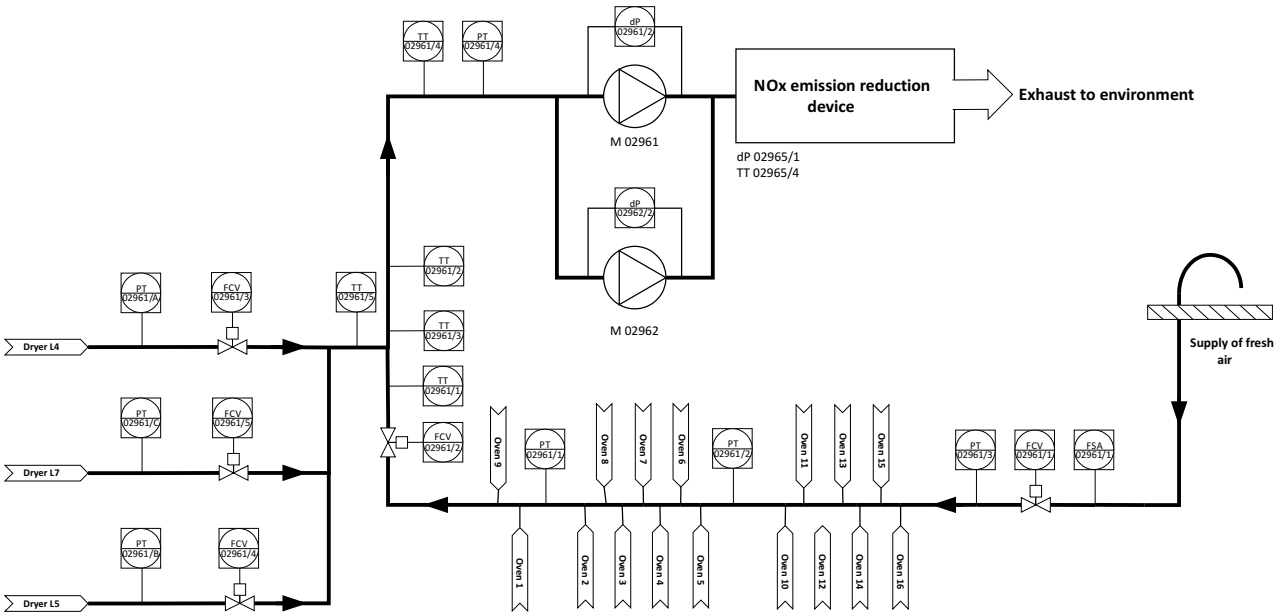


Figure 1. Air preparation process.

trol logic is programmed in accordance with a set of predefined rules based on operator expertise. The manipulated signals of dampers and ventilators are changed stepwise, with large intermediate intervals to allow the process to settle in the changed operating point. This induces substantial delays before the process adapts to any changes in the exhaust air conditions. Therefore a feasibility study was carried out to show potential improvements in reduction of emissions employing closed-loop control in the air preparation stage.

The study is based on simulation so a mathematical process model was needed. Experimentation on the actual system was not possible due to potential environmental hazard. In fact, the system involves several safety measures to prevent this and experimentation with the air preparations stage potentially induces the production shutdown, which is not acceptable.

3 Fuzzy identification

Due to nonlinear nature of the process dynamics, fuzzy logic (Zadeh, 1965) was chosen as a modelling framework and Takagi-Sugeno (TS) type model was developed. The model parameters were identified by actual production data, a separate dataset was used for model validation. The intended use of the model is to simulate a closed loop control system with some typical operation scenarios.

3.1 Takagi-Sugeno fuzzy model

With TS fuzzy model an arbitrary nonlinear system can be approximated by smooth interpolation of affine local models (Takagi and Sugeno, 1985). Every local model contributes to global model in the frame of fuzzy clusters in the space described by membership functions.

If the input data is denoted as $X = [x_1, x_2, \ldots, x_n]^T$ and the output data as $Y = [y_1, y_2, \ldots, y_n]^T$, then the model in TS form (Takagi and Sugeno, 1985) is written as a set of rules:

$$R_i : \text{if } x_k \text{ is } A_i \text{ then } \hat{y}_k = \phi_i(x_k) \quad i = 1, \ldots, c \tag{1}$$

Vector x_k represents the input data in premise while $\hat{y}_k$ is the output of the fuzzy model at time instant k. Premise vector x_k relates to fuzzy sets $(A_1, \ldots, A_c)$, where every fuzzy set A_i $(i = 1, \ldots, c)$ is characterized by a real valued membership function $\mu_{A_i}(x_k)$ or $\mu_{ik} : R \to [0,1]$ representing the membership degree of x_k with respect to fuzzy set A_i. Functions $\phi_i(.)$ are in general arbitrary smooth functions, while mostly linear or affine functions are used.

The model output in (1) can be expressed as:

$$\hat{y}_k = \frac{\sum_{i=1}^{c} \mu_{ik} \phi_i(x_k)}{\sum_{i=1}^{c} \mu_{ik}} \tag{2}$$

Equation (2) can be simplified by introducing $\beta_i(x_k)$ defined as:

$$\beta_i(x_k) = \frac{\mu_{ik}}{\sum_{i=1}^{c} \mu_{ik}}, \; i = 1, \ldots, c \tag{3}$$

In this way the degree of fulfillment of a fuzzy rule is given in a normalized form. $\sum_{i=1}^{c} \beta_i(x_k) = 1$ independently of x_k, as long as $\beta_i(x_k)$ is not zero (which can be easily assured by extending membership functions over the whole range of x_k).

Joining (2) and (3) results in:

$$\hat{y}_k = \sum_{i=1}^{c} \beta_i(x_k) \phi_i(x_k), \; k = 1, \ldots, n \tag{4}$$

The output is often defined as a linear combination of the consequence states:

$$\phi_i(x_k) = x_k \theta_i, \; i = 1, \ldots, c, \; \theta_i^T = \left[\theta_{i1}, \ldots, \theta_{i(p)}\right] \tag{5}$$

The vector of fuzzified input variables at time instant k is written as:

$$\psi_k = [\beta_1(x_k)x_k, \ldots, \beta_c(x_k)x_k], \; k = 1, \ldots, n$$

which implies fuzzified data matrix:

$$\Psi^T = \left[\psi_1^T, \psi_2^T, \ldots, \psi_n^T\right] \tag{6}$$

If the coefficient matrix for the overall set of rules is written as $\Theta^T = \left[\theta_1^T, ..., \theta_c^T\right]$, (4) can be modified to:

$$\hat{y}_k = \psi_k \Theta \tag{7}$$

and in compact form, which describes the overall data set relations:

$$\hat{Y} = \Psi\Theta \tag{8}$$

where $\hat{Y}$ is a vector of model outputs $\hat{y}_k$ and $k = 1, \ldots, n$ ($\hat{Y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n]^T$).

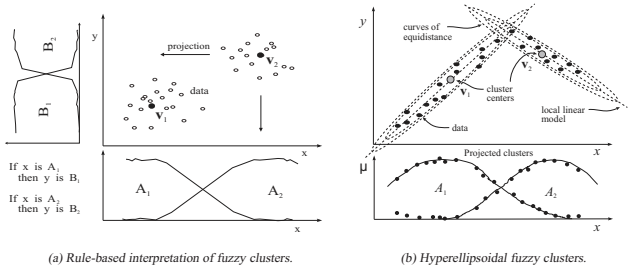
The fuzzy model (7) is often denoted as affine Takagi-Sugeno model and can be used to approximate any real valued continuous function with arbitrary precision (Kosko, 1994; Wang and Mendel, 1992; Ying and Chen, 1997). This can be proved by Stone-Weierstrass theorem. Any real valued continuous function can be approximated with a fuzzy model (Lin, 1997).

3.2 Fuzzy clustering

The TS model can be derived from available process data by identifying the structure and the parameters of the local models (Takagi and Sugeno, 1985). Structure identification includes an estimation of the cluster centers (antecedent parameters), which is usually done by fuzzy clustering. Then for each cluster the sub-model's parameters are estimated, which is usually done with a least-squares method (Chiu, 1994). Clustering can be performed in the input space only or in the product space of the input and output. The later is more general and is commonly termed shortly as clustering in the product space

The basic principle of fuzzy clustering is depicted in Fig. 2(a). Here the data is clustered into two groups with prototypes v_1 and v_2, using the Euclidean distance measure (Babuška, 1998).

If-then rules of the fuzzy model are extracted by clusters projection onto the axes. The clusters can be ellipsoids with adaptively determined shape (Gustafson and Kessel, 1979), see Fig. 2(b). From such clusters, the antecedent membership functions and the consequent parameters of the TS model can be identified (Babuška, 1998). Each obtained cluster is represented by a rule in the TS model. The principle of system identification employing fuzzy clustering consists of a series of steps depicted in Fig. 3.



(a) Rule-based interpretation of fuzzy clusters. (b) Hyperellipsoidal fuzzy clusters.

Figure 2. Identification by fuzzy clustering (Babuška, 1998).

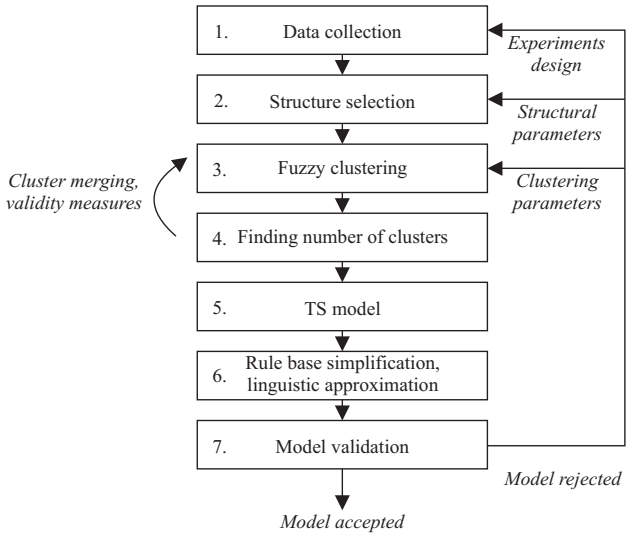


Figure 3. Identification approach based on fuzzy clustering (Babuška, 1998).

The procedure is supported by dedicated software tools, such as Fuzzy Modelling and Identification Toolbox for Matlab (FMID), which was used in the presented work. Implementation of some of the steps of Fig. 3 will be discussed in the following sections.

4 Data collection

Data collection is a key step that largely determines the quality of the modelling result. Proper identification experiments have to be designed and this is a very problematic step in several industrial processes, in particular when a process is already in operation. In some cases the process shutdown is impossible for the reason of cost or safety; in others the standalone experimentation is not possible. Because of process interconnections, it is often required that several process stages operate in line with the process under observation.

When experimentation is performed during operation, typically a number of operating restrictions have to be observed, which severely limits the experimentation possibilities.

In the presented case the air preparation process cannot run independently of main production process which delivers the polluted exhaust air. At the same time nominal operating regime has to be maintained within the main process to avoid product scrap or even automatic process shutdown, and furthermore, the emissions should stay within allowed ranges.

The first identification attempt was done with signals acquired during the control commissioning. Signals ac-

quired while the operation of control logic was tested were used as input-output data. The modelling results were not useful, because it turned out that the testing has not covered adequately large portion of the input-output space. Only manipulated values could be directly changed during testing while several other signals depend on process operating conditions that were not directly controlled. Large number of situations was simply not covered by the testing procedures, which is quite common in complex processes.

The second attempt was based on data acquired during a longer process operation testing period. The problem in this case was the built-in data compression within the data acquisition system, which writes the data only on sufficient changes. The problem was solved by interpolating available sensor measurements and by stepwise holding the manipulated values. The results were resampled by 1 *s* interval.

5 Structure selection

According to (Babuška, 1998) the structure selection in fuzzy modelling involves a set of tasks, such as: a choice of input and output variables, representation of the systems' dynamics, and a choice of the fuzzy models granularity.

5.1 Input and output variables

The control relevant output signals were specified by the customer, the same were chosen as model outputs. The inputs were chosen in accordance with the known physical background of the process. Also an expertise of the process operators was considered.

Additionally, the possibilities of data analysis with regard to determination of influential variables were tested (Glavan et al., 2013). The analysis has not brought any new insights into the input output dependencies. This indicates that for well-known processes data analysis methods are not superior to human expertise. Inputs of some of the submodels depend on the outputs of other submodels. Fig. 4 shows the initial structure of the overall process model.

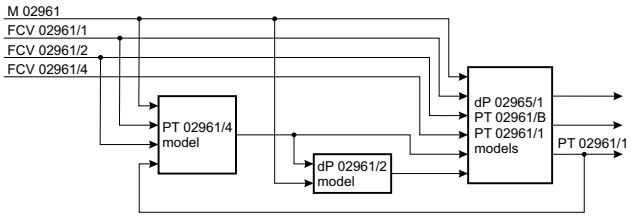


Figure 4. Initial model structure.

While individual models show good matching with the process data, the overall model of Fig. 4 exhibits large modelling error. The reason is in the error accumulation in PT 02961/1, PT 02961/4 and dP 02961/2 models (signals are labelled in accordance with process scheme in Fig. 1). Therefore, another model structure was chosen, which consists of separate models for all quantities of interest, which directly build on measured data. E.g., the model for air pressure in the ovens outlet is shown in Fig. 5.

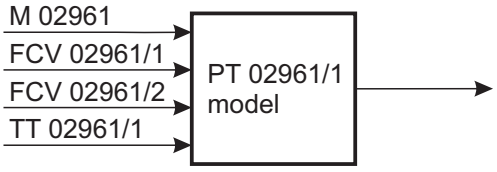


Figure 5. Fuzzy model of air pressure in the ovens outlet.

5.2 Representation of the systems' dynamics

The main design issue here was to choose the number of delayed input and output samples used as model regressors. The choice was determined by an iterative procedure, during which we started the identification with one sample delay on input and output, and then repeated the identification with increasing number of delayed outputs up to four. The effect of adding additional delayed input was also tested as well as changes of sampling interval to 2 and 3 *s*.

The quality of derived models was compared by variance accounted for (VAF) performance index (Babuška, 1998)

$$VAF = 100\% \left[1 - \frac{var(y_1 - y_2)}{var(y_1)} \right] \tag{9}$$

Among others, the above procedure led us to adjust the sampling period of the acquired data; for the further experimentation a 3 *s* sampling interval was chosen. This allowed us to use a low number of delayed signals as model regressors (2 to 3), which prevents overfitting and improves model generalization capability.

5.3 Fuzzy models granularity

When fuzzy clustering is applied to generate fuzzy models from data, the main parameter that must be chosen with regard to granularity is the number of clusters. The applied strategy here was to start with a moderate number of clusters (e.g. 5) and then experiment in a similar manner as above, gradually increasing the number of clusters up to 8, and taking VAF index as performance measure.

6 Fuzzy clustering and model validation

All models for the control relevant variables defined in Sect. II were built in a similar manner by using FMID tool and experimenting with main parameters.

Fig. 6 shows the performance of the best model for air pressure in the ovens outlet with respect to the identification signal and the validation signal.

The VAF value is 67.4 % for the identification signal and 28.9 % for the validation signal. The result of validation is rather low but the reason is in the nature of the process operation. The observed pressure largely depends

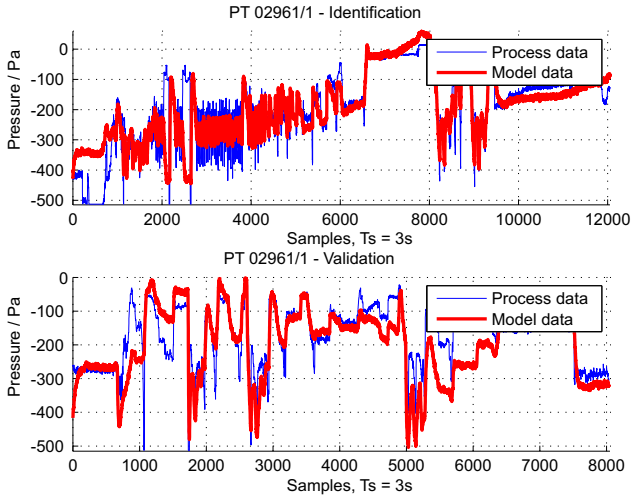


Figure 6. Performance of the fuzzy model of air pressure in the ovens outlet.

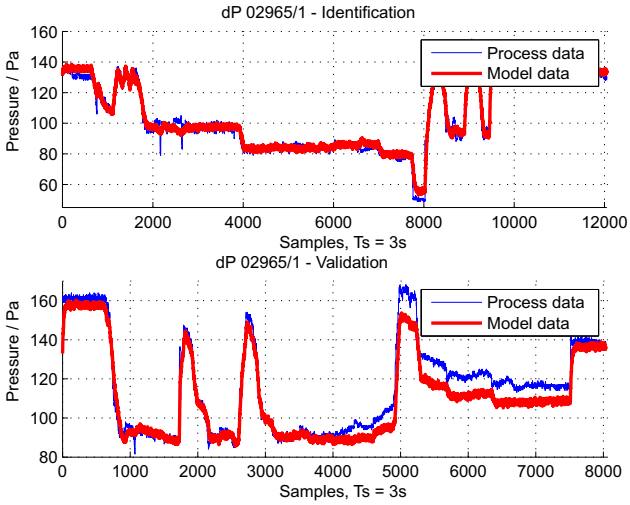


Figure 7. Performance of the fuzzy model of differential pressure in the catalyst converter unit.

on the number of the operating ovens. This parameter was not available in the process database so it could not be included as an input. It is assumed that occasional large deviations from the measured validation signals are due to change of this condition.

Fig. 7 shows a similar graph for the model of differential pressure, which is the basis for the air-flow calculation.

The VAF value is 97.8 % for the identification signal and 95.0 % for the validation signal, which is much better than in the previous case. This indicates that chosen modelling method can be very effective when proper input-output data is provided.

6.1 Comments on resulting model performance

We estimate that derived fuzzy models are of sufficient quality to test and compare various control solutions. There are noticeable deviations in responses of the PT 02961/1 model and the real system, but these are presumably mainly caused by varying conditions in the quantity of supplied exhaust air due to variable number of operating stages in the main manufacturing process.

The most relevant control variable is the air flow FIA 02965/2 (Φ_{air}) through the NOx emission reduction device, which is calculated based on dP 02965/1 ($\Delta p\ [Pa]$) and TT 02965/4 ($T\ [^oC]$) readings as follows:

$$\Phi_{air} = 3600 \cdot A \cdot v_{air} \quad [m^3/h] \tag{10}$$

$$v_{air} = \sqrt{\frac{2\Delta p}{R \cdot \rho}} \quad [m/s] \tag{11}$$

$$\rho = 1.293 \frac{273}{T + 273} \quad [kg/m^3] \tag{12}$$

where $R = 4.6285$ and $A = 2.1316\ m^2$.

As presented above, the VAF index obtained during validation shows good quality of dP 02965/1 model, which indicates the obtained model is of sufficient quality for testing air flow control strategies.

By stepwise changes of the manipulated variables the model behavior was also qualitatively evaluated, and all the signals change in accordance with experience on the real process.

Additional temperature measurements were considered as model inputs during the identification. When testing control strategy these signals are not available, therefore a set of disturbance generating models were additionally identified. Major disturbances were generated, in particular the air temperatures in different parts of the system.

These models are not very precise but enable generation of disturbances during simulated experiments with control system. Control robustness is tested this way.

7 Control experiments

As the signals observed in the real process are noisy and disturbed by additional influential variables, these aspects are added to the model used for control experiments. Noise amplitude and frequency bandwidth of the noise filter were chosen in accordance to observed process signals in the manual mode.

7.1 PID control

After model was validated and brought close to the actual process behavior by addition of noise, it was used to test various control strategies. As the focus here in this paper is not on control design, only a sample simulated scenario of the air flow is shown in Fig. 8. Note that simple PI controller was used that was tuned in a chosen operating point without noise in the model. After tuning, the control robustness was checked in the presence of noise and disturbances. Among others, additional noise cancelling measures were implemented, such as deadband on manipulated value and error filtering.

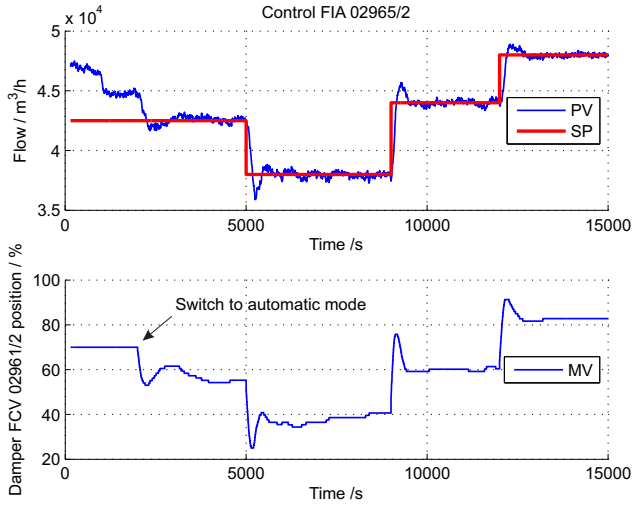


Figure 8. Illustration of the control system performance - air flow.

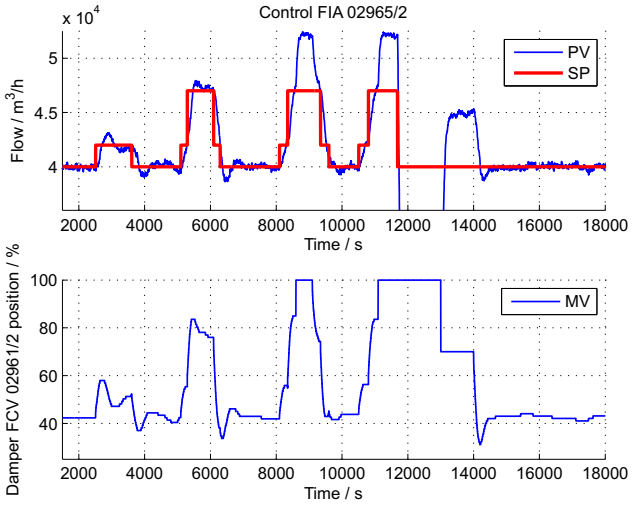


Figure 9. Illustration of the control system performance in the presence of process exceptions.

7.2 Supervisory logic

The advantage of the conventional (currently used) open-loop control system is the simple accommodation for the process exceptions. These are treated by addition of related rules in the control logic.

The main process exceptions are related to:

- start-up of the air preparation system
- exceptional process values
- invalid pressure set points.

In the following we briefly illustrate how exceptional process values are treated. The observed values are:

- high temperature in the ovens output airflow
- high temperature before the main ventilator
- high emissions concentration on the input
- high differential pressure on the main ventilator.

The desired action in all these exception is to increase the air flow through the system. In the conventional logic this is achieved by simply opening the corresponding damper for a predefined proportion.

In the proposed new control system this is accommodated by additional supervisory logic, which increases setpoint value for airflow at the predefined warning levels. With the low warning level the setpoint is increased for 2000 m^3/h, at the high warning level the setpoint is increased for 5000 m^3/h.

When lower alarm level is reached the system switches to manual mode with fully open dampers and predefined ventilator speed. At higher alarm level the system is shut down with dampers fully open.

The sample simulated operation scenario is shown in Fig. 9, where setpoint changes and related changes in the damper position can be observed, and normal system operation reestablishment can be seen when operation conditions are brought back to normal.

The results show that treatment of the exceptional conditions can be accomodated by adjustments of the setpoint values by supervisory logic.

8 Conclusions

The feasibility study shows the control of the air preparation stage could be improved with a moderate investment in the control equipment and related application software. The currently used control system is implemented in the ControlLogix 5570 family of Programmable Logic Controllers (PLCs) with PID control support within the corresponding PLC programming software. The application software could be adjusted by replacing open-loop control logic with PID controllers. Minor parameter adjustments are foreseen mainly due to different scaling of the control signals compared to simulation study.

The chosen modelling approach showed its value in relatively simple identification of the submodels when the proper model structure was determined. Nevertheless, the determination of the structure was a challenging task that was solved by using a-priori knowledge of the system behaviour in combination with experimental adjustments of the main structural parameters. Inferring this automatically from a given data set remains a difficult task, in particular in real industrial processes where possibilities of the experimentation with the input signals are limited, due to technological and safety restrictions.

The focus of the presented model development was in obtaining a model that is representative enough to enable simulated control experiments. We estimate the goal was achieved and results of the simulation experiments are satisfactory. Addition of data related to operating parameters in the main process, in particular the number of operating ovens, would open a way to improve the model and perform a feasibility study involving more advanced con-

trol strategies. Nevertheless, the potential implementation of such strategies in the industrial process under study is limited with existing control equipment.

References

R. Babuška. *Fuzzy Modeling for Control*. Kluwer Academic Publishers, Boston, USA, 1998.

Stephen L. Chiu. Fuzzy model identification based on cluster estimation. *J. Intell. Fuzzy Syst.*, 2(3):267–278, May 1994. ISSN 1064-1246.

Miha Glavan, Dejan Gradišar, Maja Atanasijević-Kunc, Stanko Strmčnik, and Gašper Mušič. Input variable selection for model-based production control and optimisation. *The International Journal of Advanced Manufacturing Technology*, 68 (9):2743–2759, 2013. ISSN 1433-3015. doi:10.1007/s00170-013-4840-1.

D. E. Gustafson and W. C. Kessel. Fuzzy clustering with a fuzzy covariance matrix. In *Proc. 18th IEEE Conf. Decision and Control*, pages 761–766, Jan 1979.

Bart Kosko. Fuzzy systems as universal approximators. *IEEE Trans. Comput.*, 43(11):1329–1333, November 1994. ISSN 0018-9340. doi:10.1109/12.324566.

Cheng-Jian Lin. SISO nonlinear system identification using a fuzzy-neural hybrid system. *Int. J. Neural Systems*, 8(3):325–337, 1997. doi:10.1142/S0129065797000331.

T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. Systems, Man, and Cybernetics*, SMC-15(1):116–132, Jan 1985. ISSN 0018-9472. doi:10.1109/TSMC.1985.6313399.

L. X. Wang and J. M. Mendel. Fuzzy basis functions, universal approximation, and orthogonal least-squares learning. *IEEE Trans. Neural Networks*, 3(5):807–814, Sep 1992. ISSN 1045-9227.

Hao Ying and Guanrong Chen. Necessary conditions for some typical fuzzy systems as universal approximators. *Automatica*, 33(7):1333–1338, 1997. ISSN 0005-1098.

L.A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965. ISSN 0019-9958.

From Iterative Balance Models to Directly Calculating Explicit Models for Real-time Process Optimization and Scheduling

Tomas Björkqvist[1] Olli Suominen[1] Matti Vilkko[1] Mikko Korpi[2]

[1]Department of Automation Science and Engineering, Tampere University of Technology, Finland, {tomas.bjorkqvist,olli.suominen,matti.vilkko}@tut.fi

[2]Research Center Pori, Outotec Oy, Finland, mikko.korpi@outotec.com

Abstract

Optimal utilization of complex processes involves real-time operational optimization and scheduling, especially in cases where the production line consists of both continuous and batch operated unit processes. This kind of real-time optimization requires process models which can be computed significantly faster than real-time. Iterative balance calculation is typically far too slow for these cases. This paper presents a method for converting an iterative balance model to a directly calculating model suitable for on-line process optimization. The approach is demonstrated with the first unit process in the copper smelting line, the flash smelting furnace (FSF). The method consisted of formulating an equation group based on the constrained FSF HSC-Sim model and solving the unknown parameters and static states with use of a symbolic calculation software. The solution was implemented as a function whose calculation time fulfilled the requirements for scheduling use.

Keywords: real-time model, static, mass balance, equation group, symbolic computation, metallurgy, copper smelting, scheduling

1 Introduction

The general digitalization of society has brought on a pronounced digitalization wave in process industry. The benefits of digitalization are not fully utilized in some conventional industrial processes and there new advantages are available which can improve their efficiency and ability to stay competitive in increasing global competition. Often the design of these industrial processes is based on long term empirical and theoretical knowledge which has been incorporated into thoroughly built mathematical models. These models often include iterative balance calculations to fulfill empirical and physical process constrains. These models are well suited for steady state process design and often used when offering, planning and constructing new process lines.

Optimal utilization of processes typically involves real-time operational optimization and scheduling, especially in cases where the production line consists of both continuous and batch operated unit processes. This kind of real-time optimization requires process models which can be computed significantly faster than real-time. Iterative balance calculation is often far too slow for these cases. The high demand on execution time can often be compensated by lowering demands on model precision for the real-time operation optimization. Examples of demanding real-time optimization utilized in process design can be found in (Harjunkoski et al., 2016; Touretzky et al., 2016; Pelusi 2012a; Pelusi 2012b).

Good examples of thoroughly built steady state models can be found in metallurgy. Most metallurgical processes are old and have large societal impact which has allowed extensive development work to model process behavior over many decades. These processes comprise complex physical and chemical reactions and modelling has been both theoretical and empirical. To fulfill the basic requirement of mass and energy conservation and empirical observations iterative calculation is often employed.

The incentive for this study is the need for operational optimization of a copper smelting line. Optimal operation of a copper smelting line is challenging for the operators as the operation is divided into many complex individual sub processes. Plant wide operation is required to maximize production and resource efficiency. Additionally, more challenging ores have to be used to retain economic competitiveness worldwide which increases the need for process optimization. Improved operation of copper smelting can provide improved utilization of different input materials and recyclants. Copper smelters present a challenging optimization problem where the harsh environment can prevent obtaining mineral and operational information, data is highly uncertain or measurements may be severely delayed. A full scale optimization of the complete process line will include a considerable amount of variables and require the consideration of large time horizons. Further, many of the underlying models are nonlinear. Thus, sub processes and the related models should be relatively lightweight in terms of their computational requirements. In principle, the development of optimization for a copper smelting line

operation consists of modelling of unit processes and designing of optimization / scheduling for the combined unit process models.

Static input output process models can be derived with use of mass and energy balances supplemented with sometimes uncertain process reaction knowledge completed with empirical knowledge. In principle this empirical knowledge can be written as constraints in equation form. These equations can be completed with mass and energy balances to form a complete equation group determining process reactions. By solving the equation group, the unknown parameters and thereby the static process state can be solved under the given constraints. In practice this approach is challenging as the equations are often complex and manual solutions may be error prone and exceptionally time consuming. Development of aids for this challenge started in the beginning of the 1970s under the scientific area of symbolic computation. Software programs for manual computation are called computer algebra systems (CAS) and are at present highly developed and even implemented in hand held calculators. These systems include Mathematica (Wolfram) and Maple (Maplesoft), the latter has been implemented in Matlab (Mathworks) as the Symbolic Math Toolbox. In later Matlab versions, the toolbox is based on the MuPAD symbolic engine originally developed at the University of Paderborn. Matlab offers a convenient way of shifting from symbolic calculus to numeric powerful computation.

Utilization of symbolic computation for solving unknown variables of restricted mass balance equations seems to be a rare approach or rarely reported. A similar method was used in (Korpela et al., 2014) in the same research group but the authors have not found similar work by others. Symbolic computation is, however, commonly utilized when forming first principle models (Belkhir et al., 2015; Lin et al., 2009; Yakhno et al., 2016). Its use is especially convenient for model design with e.g. Lagrangian mechanics (Moosavian et al., 2004).

For optimization of the operation of the copper smelting line computationally lightweight models of all unit processes are required. This paper presents a method for converting an iterative balance model to a directly calculated model suitable for process operation optimization. The method is demonstrated with the first unit process in the copper smelting line, the flash smelting furnace (FSF).

2 Copper Production Line

Copper smelting begins from the mixing of a suitable concentrate mix with a copper content of 20-30 % which, after drying, is fed to the FSF. The mix reacts with the oxygen-enriched air feed and separates to matte (~62-70 % Cu) and slag. These are removed intermittently from the FSF, matte is moved to the converters, and slag is processed further in the slag treatment plant. After treatment, both FSF and converter slag can be recycled back to the FSF. The matte copper content can be viewed as one of the main decision variables in smelting as the higher copper content in matte is, the higher the copper content in the slag. Additionally, it is often used as a variable in separation of other valuable metals to both matte and slag. Silica flux is added to the FSF feed and to converters during operation to achieve suitable conditions for separation of matte and slag. One of the main bottlenecks for operation is the capacity of the gas treatment plant which produces sulphuric acid from the off gasses of both the FSF and converters.

Pierce-Smith converters use a submerged feed of oxygen enriched air. Converters are operated in batches where first, in multiple slag-making stages, FSF matte is added between air blows. Here, most of the iron compounds will react and move to slag. Second, in one longer copper-making stage the remaining sulphur is removed from copper compounds. Temperature is controlled with the addition of recycled material, e.g. scrap metal. Finally, the ensuing blister copper (~99 % Cu) is moved to anode furnaces where oxygen is removed from the matte and copper is cast to anodes for transportation to electrolysis. Figure 1 shows a full copper production line including both smelting and refining. A detailed description of the smelting process can be found for example in (Schlesinger et al., 2011).

3 Model Conversion

The method for converting an iterative balance model to a directly calculating model is here demonstrated with a model of the flash smelting furnace, modelled in HSC-Sim (Outotec). HSC-Sim is a calculation module of HSC Chemistry software developed by Outotec. The name refers to the automatically utilized thermochemical database which contains enthalpy (H), entropy (S) and heat capacity (Cp) data for an extensive amount of chemical compounds. The HSC-Sim module enables application of HSC Chemistry to a whole process made up of process units and streams. The HSC-Sim module consists of a graphical flowsheet and spreadsheet type process unit models. The custom-made variable list enables creation of different types of process models in chemistry, metallurgy, mineralogy, economics, etc. Each process unit is actually one Excel file. In the Distribution units the compounds are divided into elements and calculation is done with element distribution coefficients. Based on process knowledge some coefficients are defined as fixed. Coefficients for assisting elements in compound formation are calculated based on molar need and supply and called float. Surplus elements are divided with coefficients called rest. Units can be used together or separately and the calculations can be Excel- or DLL-based.

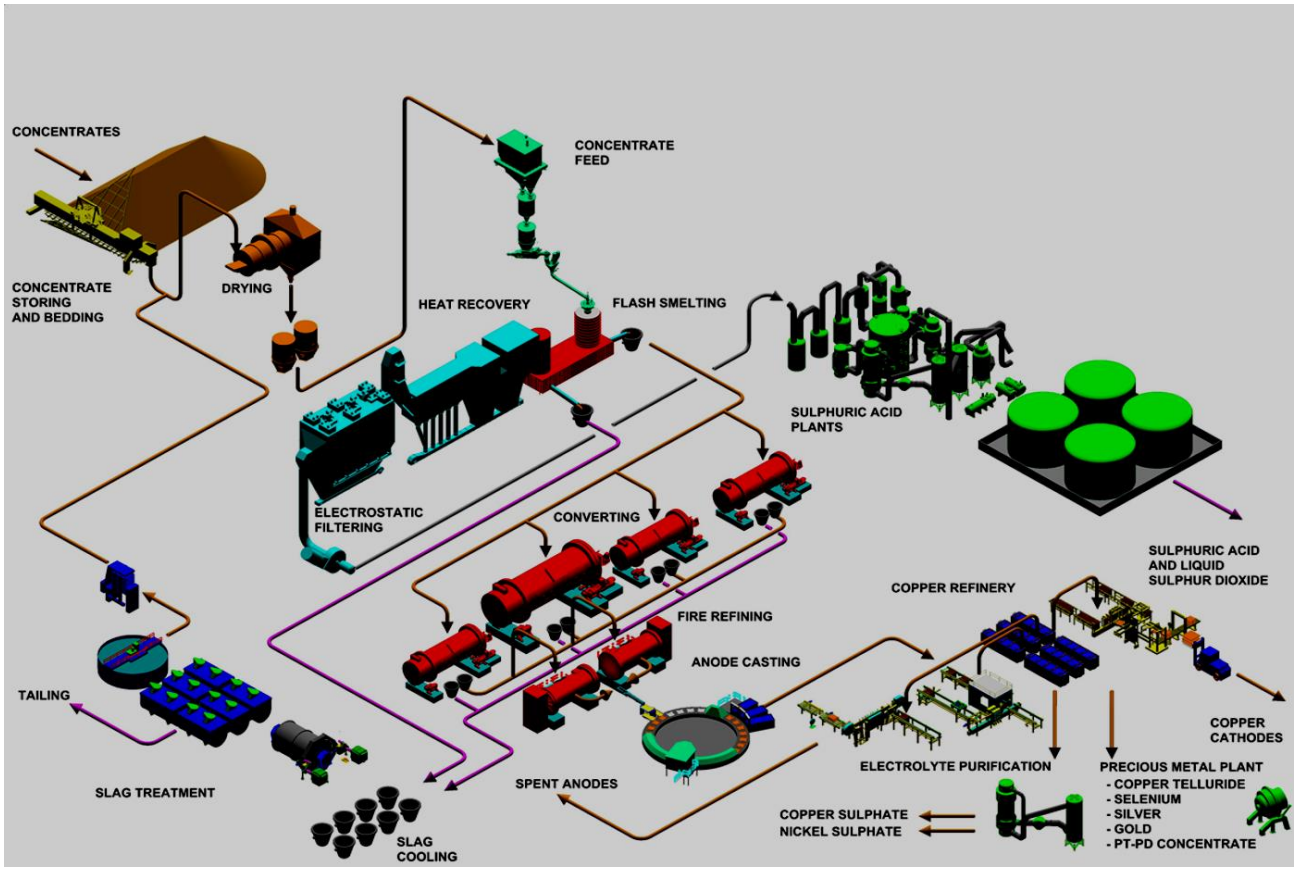


Figure 1. Flow sheet of copper process at Boliden Harjavalta (Boliden).

HSC Sim pyro models are mathematical process models based on mass- and energy balances and empirical knowledge controlling the equilibrium state. These models are successfully used in strategic planning of metal processing. The drawback of these models is the iterative calculation needed for reaching the equilibrium state. This iterative calculation is too slow for use in on-line process optimization.

3.1 Legacy model

The flash smelting furnace process has been modelled in HSC-Sim as a static division process with empirical knowledge controlling parts of the division coefficients. The implementation is a spreadsheet-like division calculation with iterative calculation to fulfill constraints derived from empirical and physical knowledge.

The model consists of three main spreadsheets; Input, Distributions and Output, each containing 146-424 rows and 68 columns. The Input sheet is sparsely filled with element mass flows and describes how input compounds in different streams are broken up to elements according to chemical molar consistency. The Distributions sheet is sparsely filled with distribution coefficients dividing element mass flow into compounds for different output streams partly according to chemical reactions. The Output sheet is filled with corresponding element mass flows that build up the output compounds in different output streams. Additionally, to the three main spread sheets, a Controls sheet includes 27 empirical process observations that must be fulfilled in the stationary state.

In principle, the distribution from input compounds to output compound is built up around how the main elements copper (Cu) and iron (Fe) is distributed between compounds in the output streams. The chemical reactions requre assisting element as oxygen (O) and silicon (Si) which are brought in as floating elements. Sulphur (S) is partly handled as a main element and partly as an assisting element. As a result the model consists of some fixed distribution coefficients, many coefficients which are iteratively adjusted to fulfill the empirical observations and numerous coefficients calculated as float according to corresponding chemical reactions or as rest for surplus elements. The model is thus a system of four spread-sheets with a large number of interconnected cells. An iterative routine is used to solve the distribution coefficients and thereby the element and compound streams in the stationary state.

The calculation is very useful for off-line strategic planning of metal processing. The calculation is, however, too slow for real-time process optimization.

3.2 Method for derivation of fast calculating model

In general, the objective for the study was to find a method for converting iterative output controlled balance models to directly calculating models suitable for process scheduling. The basic idea was to form a symbolic equation group based on the flash smelting furnace HSC-Sim model and to solve this group analytically with symbolic computation to achieve causal outputs as direct functions of inputs. The solution is possible due to empirical knowledge included in the Controls sheet of the FSF HSC-Sim model.

Thus, the task was to write a fully parametrized equation group based on the FSF HSC-Sim model where the equations are based on the equations of empirical knowledge in the Controls sheet. The model is in this analytic approach simplified. The input elements include only the main elements; copper (Cu), iron (Fe), nitrogen (N), oxygen (O), sulphur (S), silicon (Si) and other content (Ot). The distribution of the elements between the output streams, which are settler gas, settler fume, settler dust, slag and matte, is fully in line with the FSF HSC-Sim model. The eight equations determining empirical knowledge regarding the main elements was chosen as base for the equations. To enable an analytic solution with the symbolic software the equation group has to be exactly determined.

The equation group formulation starts with defining all basic variables as symbolic variables. This example included 7 element mass flows, 23 distribution coefficients for element distribution to output streams and 41 distribution coefficients for element distribution into compounds in the different output streams. The main formulation work is to define the relationship between these variables with emphasis on the formulation of the float and rest variables. Here, this part required about 75 definitions. After these definitions, the output compounds can be formulated. Afterwards, the final equations based on the empirical knowledge in the Controls sheets can be written. To ease the derivation of the analytic solution of the software the nonlinearities in the empirical knowledge were linearized. The same variables as the manipulated variables in the iterative solution of HSC-Sim model were chosen as variables for the calculation to solve. They were; distribution coefficient for Fe to matte, distribution coefficient for Fe in slag to FeS, distribution coefficient for Cu to slag, distribution coefficient for Fe in matte to Fe_3O_4, Ot to matte, Si input stream, O input stream and distribution coefficient for Fe in slag to Fe_3O_4.

This study utilizes the Symbolic Math Toolbox in the Matlab software. With the relationships concerning use of oxygen still undefined, the solver managed to achieve a fully symbolical solution in around five minutes with a laptop. When oxygen is taken into account, the solver has been forced to settle for a numeric approximation, which still includes all the variables in an appropriate manner. The length of the analytic solutions is over 25 000 characters. The solutions are at this stage provided with the values of the fixed variables. The last task of the program is to produce usable functions of the long analytic solutions.

4 Model Validation and Discussion

Model validation is performed to ensure usability of the model in real-time process optimization and scheduling. As copper content in matte is a good measure of the process state, the validation is performed at varying matte copper percentage.

4.1 Similarity to legacy model

Figure 2 shows a comparison between the analytical direct solution results, with the blue line, and iteratively calculated HSC-Sim results, red line, as function of matte copper percentage.

The cause for the differences is the fact that the analytically solved model is a simplified model of the process including only the main elements. E.g. both silicon and oxygen is consumed by other minor compounds which are not included in the model. The difference is mainly a shift of magnitude which can easily be compensated by a term proportional to the total concentrate flow. With this compensation the analytically solved model is adequate for the on-line utilization.

4.2 Calculation time

As the optimization and scheduling algorithm calls the model hundreds of times per second the calculation time has to be short. A test function call from Matlab showed that the execution time is only some milliseconds for calls of two to eight variables, which is sufficient for the on-line utilization. The calculation time for the iterative solution of the HSC-Sim model is tens of seconds.

5 Model Utilization

The directly calculating model of the flash smelting furnace process will be utilized in scheduling of a copper production line to optimize production and costs. When solving the equation group the solvable variables can be freely chosen. There are two evident ways of model formulation that can be utilized.

5.1 Direct input output

A natural solution would be to form a direct input output model to mimic the real smelting process. Figure 3 represents a scheduling structure that utilizes the input output model. As scheduling is a high level task whose interests are in production rate and oxidation level in first stage smelting, a lower level control structure has to deal with the unit control of the flash smelting furnace. This is shown as feedback control of the open

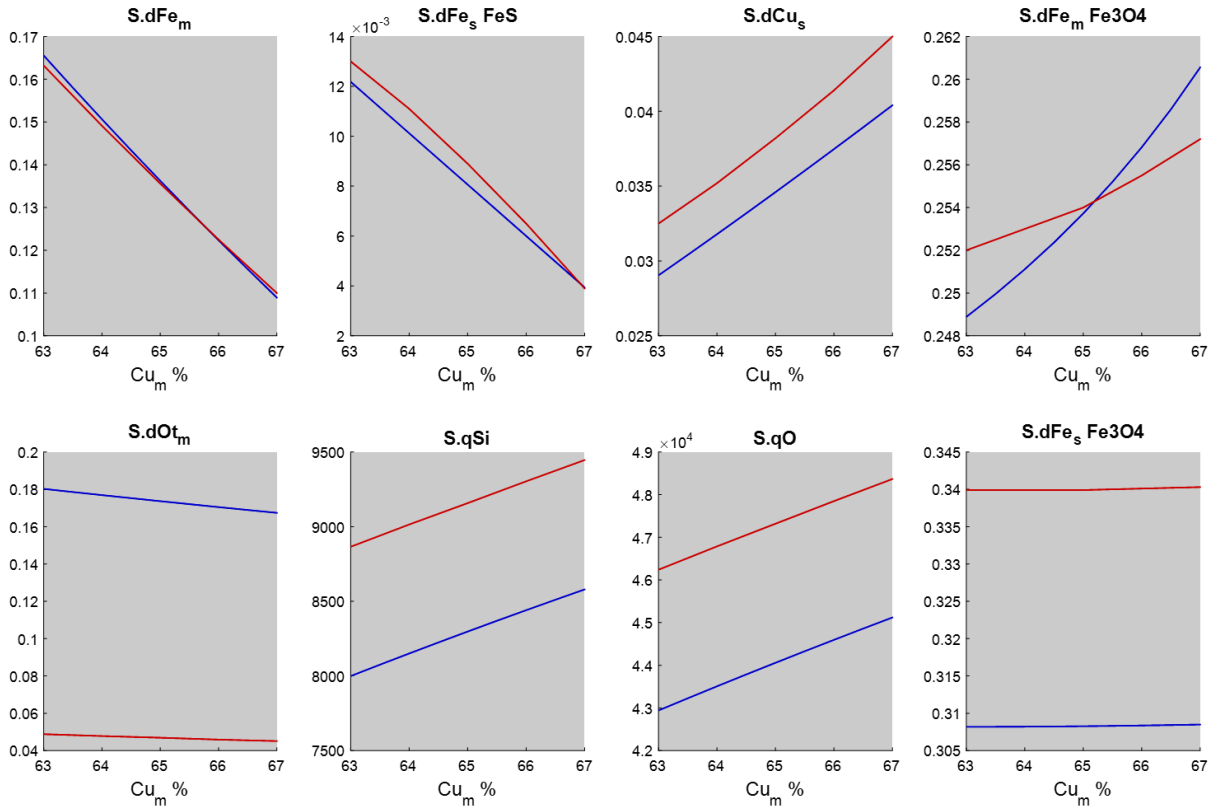


Figure 2. A comparison between analytical solution results with blue line and iteratively calculated HSC-Sim results with red line.

loop model. In practice, this could be a sub optimization task for the scheduling routine.

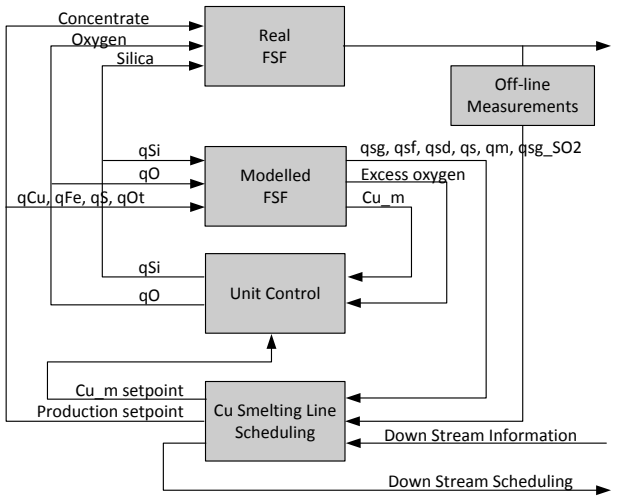


Figure 3. Direct input output model utilized in scheduling.

5.2 Closed analytic solution

To enhance the direct scheduling interests, the required control variables can directly be chosen as solvable variables in the equation group. The static model allows us to utilize a closed analytic solution whose scheduling structure is clear and shown in Figure 4. This direct solution will not need the sub optimization. Feedback from the off-line measurements compensates for model inaccuracy.

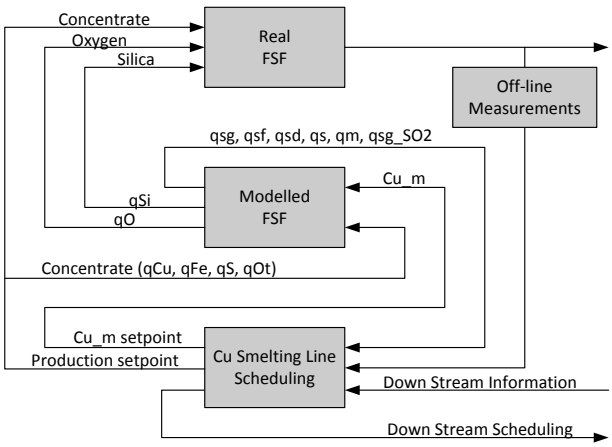


Figure 4. Closed analytic solution utilized in scheduling.

6 Conclusions

The objective of this study was to develop a method for converting iterative output controlled balance models to directly calculating models for process optimization and

scheduling. This method was used in the case of a flash smelting furnace, previously modelled in HSC-Sim. The fast calculating model is to be used in optimization of the total production line operation.

The method consisted of formulating an equation group based on the constrained FSF HSC-Sim model and solving the unknown parameters and static states with use of a symbolic calculation software. The study was successful even if it requires careful formulation work and the solution matched the solution of the original model. The equation group should be fully determined to enable a solution. The solution was implemented as a direct calculation function whose calculation time fulfilled the requirements for scheduling use.

The advantage with the approach is that even though the length of the generated functions disables model maintenance in function form, functions can easily be recalculated after updates in the HSC-Sim model are done. The modelling method has shown to be a powerful general way of converting complex iteratively solvable models to fast directly calculating models for utilization in process optimization and different operator advisory systems.

The presented demonstration model did not include an energy balance and thereby the amount of nitrogen (N) feed is kept constant even if the nitrogen feed is in practice the means to affect process temperature. The legacy model is built on the assumption that temperature is on normal level which enables a mass balance without temperature dependency. The energy balance will be included in future work.

Acknowledgements

This work was carried out in the SIMP research program coordinated by the Finnish Metals and Engineering Competence Cluster (FIMECC) Ltd. The support is gratefully acknowledged. The authors are also grateful for the process expert knowledge and good working possibilities provided by Petri Latostenmaa and Ville Naakka at Boliden Harjavalta.

References

W. Belkhir, N. Ratier, D. D. Nguyen, B. Yang, M. Lenczner, F. Zamkotsian and H. Cirstea. Towards an automatic tool for multi-scale model derivation illustrated with a micro-mirror array. *In 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing 2015*, pages 47-54. doi 10.1109/synasc.2015.17.

Boliden. Available via www.boliden.com/Documents/Press/Publications/Place%20broschures/boliden-harjavalta-en.pdf. [accessed June 23, 2016].

I. Harjunkoski and R. Bauer. Configurable Scheduling Solution using Flexible Heuristics. *In Proceedings of the 26th European Symposium on Computer Aided Process Engineering – ESCAPE 26 2016*, pages 2362-2366.

T. Korpela, T. Björkqvist, Y. Majanne and P. Lautala. Online Monitoring of Flue Gas Emissions in Power Plants having Multiple Fuels. *IFAC-PapersOnLine*, 19(1): 1355-1360, 2014. doi: 10.3182/20140824-6-ZA-1003.01913.

J. Lin and C. Chen. Computer-aided-symbolic dynamic modeling for Stewart-platform manipulator. *Robotica*, 27: 331-341, 2009.

Maplesoft. Available via www.maplesoft.com/solutions/education/. [accessed June 23, 2016].

Mathworks. Available via www.mathworks.com/. [accessed June 23, 2016].

S. A. A. Moosavian and E. Papadopoulos. Explicit dynamics of space free-flyers with multiple manipulators via SPACEMAPLE. *Adv. Rob.*, 18: 223-244, 2004.

Outotec. Available via www.outotec.com/en/Products--services/HSC-Chemistry/Calculation-modules/Sim--process-simulation/. [accessed June 23, 2016].

D. Pelusi. PID and intelligent controllers for optimal timing performances of industrial actuators. *International Journal of Simulation: Systems, Science and Technology*, 13(2): 65-71, 2012.

D. Pelusi. Improving settling and rise times of controllers via intelligent algorithms (2012). *In Proceedings – 2012, 14th International Conference on Modelling and Simulation, UKSim 2012*, art. no. 6205447, pages 187-192.

M. E. Schlesinger, M. J. King, K. C. Sole and W. G. Davenport (2011). *Extractive metallurgy of copper. Elsevier*.

C. Touretzky, I. Harjunkoski and M. Baldea. A Framework for Integrated Scheduling and Control using Discrete-Time Dynamic Process Models. *In Proceedings of the 26th European Symposium on Computer Aided Process Engineering – ESCAPE 26 2016*, pages 601-606.

Wolfram. Available via www.wolfram.com/mathematica/. [accessed June 23, 2016].

V. Yakhno and M. Altunkaynak. A polynomial approach to determine the time-dependent electric and magnetic fields in anisotropic materials by symbolic computations. *COMPEL - the International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, 35: 1179-1202, 2016.

Principal Component Analysis Applied to CO_2 Absorption by Propylene Oxide and Amines

M. H. Wathsala N. Jinadasa[1], Klaus J-Jens[1], Carlos F. Pfeiffer[1], Sara Ronasi[2], Carlos Barreto Soler[2], Maths Halstensen[1]

[1]Faculty of Technology, Natural Sciences and Maritime Sciences – University College of Southeast Norway, Post box 235, N-3603 Kongsberg, Norway, {Wathsala.jinadasa, Klaus.J.Jens, carlos.pfeiffer, Maths.Halstensen}@usn.no

[2]Norner Research, Asdalstrand 291, N-3962, Stathelle, Norway, {sara.ronasi, carlos.barreto}@norner.no

Abstract

Carbon dioxide absorption by mixtures of propylene oxide / polypropylene carbonate at 60^0C was monitored by Raman spectroscopy at 20, 40 and 60 bar in a 2 L autoclave reactor. Multivariate preprocessing techniques were used to process raw Raman spectra and Principal Component Analysis was performed. Simulation data from the Peng- Robinson equation of state were used to model the absorbed CO_2 amount and spectroscopic signals. Results showed that Principal Component Analysis can be used to explore the dynamics of the system at different pressure levels and to track the CO_2 absorption. A similar analysis was carried out to monitor CO_2 absorption by four different amines at room temperature and pressure in a batch reactors. The CO_2 content was determined from titration and was used to model the spectroscopic data. Principal Component Analysis proved to be able to identify CO_2 absorption capacity in the amines. This feasibility study confirms that Raman spectroscopy together with multivariate analysis can effectively report chemical information and dynamics in these CO_2 absorption systems and hence can be used for developing regression models for online monitoring and control.

Keywords: principal component analysis, CO_2 absorption, propylene oxide, amines

1 Introduction

Carbon dioxide (CO_2) is known to be the primary greenhouse gas contributing more than 60% of global warming. Capturing CO_2 from power plants and industrial sources and utilization them to produce usable products is of paramount importance from a standpoint of "waste to money". Absorption of CO_2 by amines is one of the most popular technologies for CO_2 capture. Amines are categorized as primary, secondary or tertiary amine based on their chemical structure. The reaction between amines and CO_2 is complex (McCann *et al*, 2009). However, when considering the CO_2 mass balance, it can be seen that once absorbed by a primary amine, CO_2 will remain in the form of carbonate, bicarbonate, carbamate or molecular CO_2 as given in (1). When it is a tertiary amine, there is no carbamate formation (2).

Synthesis of polypropylene carbonate (PPC) by reaction of CO_2 and propylene oxide (PO) in the presence of a catalyst has become a fascinating research area as a CO_2 utilization technique to produce a polymer out of a waste greenhouse gas (Jiang *et al*, 2014). In the presence of a catalyst, the chemical reaction of PPC synthesis takes place as given in (3).

CO_2 absorption capacity by an amine or by in the liquid phase PO is a key performance criteria in industrial scale CO_2 capture and polymerization processes. However, the measurement of CO_2 absorption in these mixtures are challenging and require proper understanding of the chemistry behind reaction (1), (2) and (3). Several offline analytical instruments and chemical methods are available such as titration, Nuclear Magnetic Resonance spectroscopy and gas chromatography to determine the CO_2 absorption in both applications above. Most of these methods are time consuming. A fast, online method to detect CO_2 absorption is important in process monitoring and control. Considering the in-situ performance, Raman spectroscopy can be suggested as a competitive approach for this purpose. It gives chemical information of a sample as a function of Raman wavenumber and scattered light intensity. When converting the information given by a Raman spectroscopy, multivariate calibration is required to transform the spectroscopic measurement into informative output. Raman spectra contain several wavenumbers or group of wavenumbers which are chemically important and needed to be included in the multivariate regression models. However, it is often misleading to use

traditional multilinear methods such as ordinary least square for calibration, when a single wavenumber (X variable) is not sufficient to predict the useful quantity (y variable); when X variables are highly correlated or when there is no adequate information to understand which X variables are correlated to the y variable. In such instances, multivariate analysis gives the advantage of overcoming the collinearity problems while preserving useful information hidden in collinear data. In this study, Principal Component Analysis (PCA) which is a fundamental multivariate analysis tool, has been used as a data compression and exploratory method to investigate the feasibility of Raman spectroscopy as a viable analytical technology to quantify CO_2 absorption by amines and propylene oxide. Eight experimental cases have been used in this analysis. Four of them are related to CO_2 absorption by PO and a mixture of PO and PPC. These experiments were meant to compare CO_2 absorption in the CO_2-PO system with respect to the CO_2-PO-PPC system at some selected process conditions. The other 4 experiments were used to identify CO_2 uptake by four liquid amine solvents. These solvents are currently in research interest to capture CO_2 from flue gas in power plants and industries (Leung *et al*, 2014).

2 Methods

Experimental description of 8 test cases are presented in Table 1. Six organic chemicals were used in the experiments and they are given in Table 2. Case 1-4 were carried out in a closed 2L steam jacketed autoclave reactor equipped with a stirrer while the pressure was increased gradually by adding CO_2 to the reactor. Case

Table 1: Description of test cases.

Case Number	CO_2 loaded solution	Description
1	PO in non-stirred condition	Each case has one sample in a 2L reactor at 60^0C. Tested pressure levels :20, 40 and 60 bar Stirrer speed = 400 rpm
2	PO in stirred condition	
3	PO+PPC in non-stirred condition	
4	PO+ PPC in stirred condition	
5	MEA 37 samples	Each sample in 10 mL glass reactor. Reaction between CO_2 and amine took place at room temperature and pressure
6	3AP 42 samples	
7	3DMA1P 41 samples	
8	MDEA 41 samples	

1 and 2 were PO-CO_2 binary mixtures while Case 3 and 4 were PO-PPC-CO_2 ternary mixtures. A Raman immersion probe was connected through the bottom of the reactor and signals were acquired continuously with time. In case 5-8, CO_2 absorption on liquid amines was observed under equilibrium condition at room temperature and pressure. Raman signals were recorded by immersing the Raman probe into sample reactors after allowing each sample to reach equilibrium.

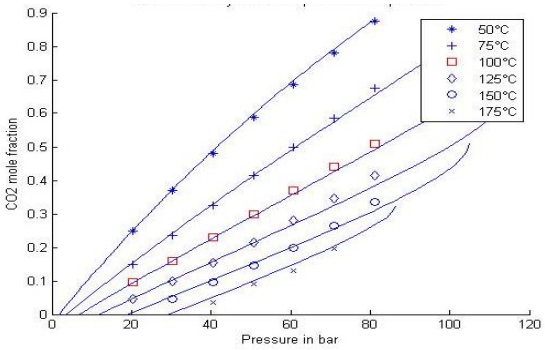
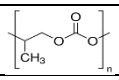
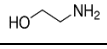
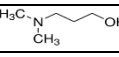
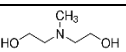


Figure 1. CO_2 mole fraction of PO-CO_2 system at different pressures and temperatures (Peng-Robinson model with binary interaction parameter equal to 0.281).

Table 2. Description of materials.

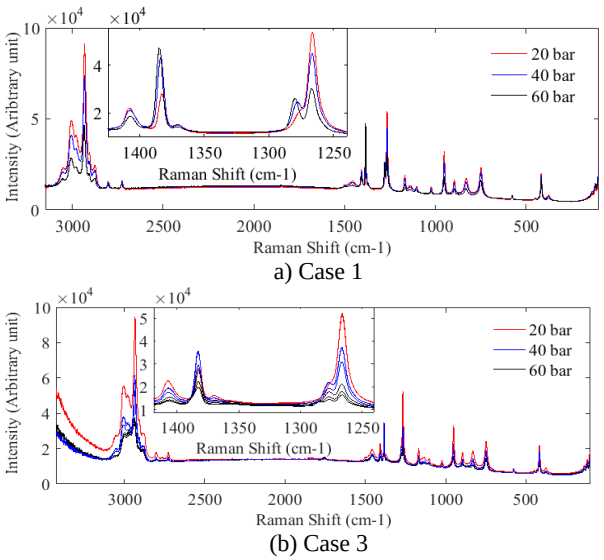
Name	Abbreviation	Chemical structure	Chemical category
Propylene oxide	PO		epoxide
Polypropylene carbonate	PPC		a copolymer of CO2 and PO
2- Aminoethanol	MEA		Primary amine
3-Amino-1-propanol	3-AP	$H_2NCH_2CH_2CH_2OH$	Primary amine
3-dimethylamino-1-propanol	3DMA1P		Tertiary amine
Methyl diethanolamine	MDEA		tertiary amine

2.1 CO_2 in polymer solutions – from thermodynamic models

In this study, Raman signals (X variables), were calibrated with the absorbed CO_2 content (y variable). Reliable measurement of y variable in Case 1-4 using an analytical method is challenging as CO_2 quickly desorbs if a sample is taken out from the reactor for analysis. Therefore, the CO_2 content data at required pressure and temperature were calculated from the vapour-liquid equilibrium (VLE) data of CO_2-PO system generated using the Peng-Robinson equation of state. The Peng-Robinson model was fitted using experimental data reported in (Chen *et al*, 1994; Shakhova *et al*, 1973). Figure 1 shows predictions of the CO_2 mole fraction in PO-CO_2 system using Peng-Robinson model simulated in Aspen Plus V7.2 software which shows that the absorption of CO_2 at a constant temperature gives a linear behavior with pressure. This linear relationship was taken to model the CO_2 mole fraction at 60^0C at which the experimental cases of 1-4 were carried out.

2.2 CO_2 in amine solutions – from titrations

In experiments from case 5-8, each sample contained 30 % of solvent (solvent weight/total weight of water and solvent) but different amounts of CO_2 added. They were prepared in 10 mL glass reactors and after reaching equilibrium a titration method was carried out to measure its true CO_2 content in units of moles CO_2 per mole solvent.



a) Case 1

(b) Case 3

Figure 2. Raman signals of CO_2 loaded polymer samples.

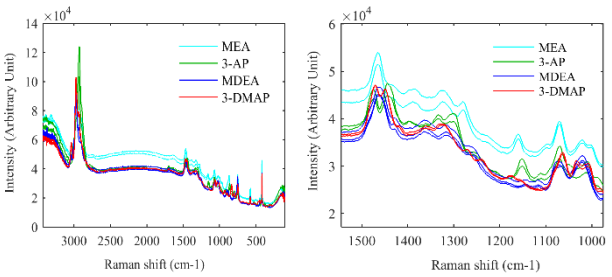
2.3 Raman Spectroscopy

Raman spectroscopy used in this study was Kaiser RXN2 Analyser of 785 nm laser wavelength, 400 mW laser power and 100-3425 cm^{-1} spectral range. An immersion optic probe which is connected to the RXN2 Analyser via a fibre optic cable, carries the laser light to the sample and in-elastically scattered Raman light is conveyed back to the instrument. The instrument output is a plot of intensity of scattered light versus energy difference (given by wavenumber in cm-1) which is called a Raman spectrum. Peaks and their intensity in a Raman spectrum carry information about the chemicals and their composition respectively.

2.4 Data Processing

For a set of n objects (eg: different samples or signals with time), a Raman spectroscopy measurement generates a data matrix of n x p where p is 3326 Raman wavenumbers. This data matrix contains useful information about the chemical fingerprint of objects as well as noise. They are also called residuals which can be due to the interference of other chemical components, laser input variations or instrument noise. Unless any data conditioning method is applied to remove this unwanted structure from the data matrix, calibration of spectroscopic signals will not be reliable and do not really generate a model which really represent the variation of analyte of interest.

Three data pre-processing techniques were applied for raw Raman data. These were baseline-whittaker filter, standard normal variate (SNV) and mean centering. The baseline-whittaker filter available in PLS toolbox in Matlab is an extended version of (Eilers, 2003) where a weighted least square method is applied to remove background noise and baseline variations. A detailed description of the algorithm can be found in the original work (Eilers, 2003) and (Atzberger *et al*, 2010). Some



(a) 100-3325 cm^{-1} (b) 950-1550 cm^{-1}

Figure 3. Raw spectra of CO_2 loaded samples (Case 5-8).

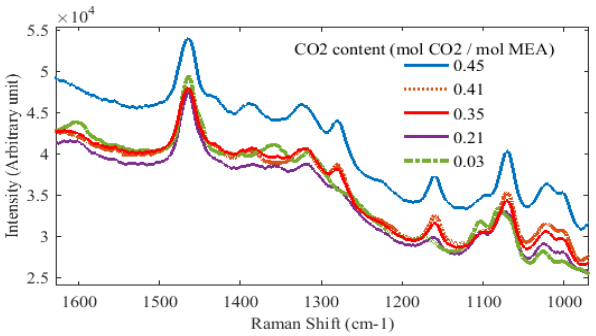


Figure 4. Raw spectra of CO2 loaded MEA samples (Case 5).

spectra which should be otherwise identical, become different due to baseline and pathlength changes. SNV was applied to remove these scatter effects in the spectra which were specially observed in case 1-4. The algorithm is similar to autoscaling row wise and hence corrects each spectrum individually (Barnes *et al*, 1989). By mean centering of data, each column in the data matrix is centered by subtracting the mean. It is reported that by mean centering, rank of the model is reduced, data fitting accuracy is increased and offset is removed (Bro *et al*, 2013).

2.5 Principal Component Analysis (PCA)

Principal component analysis is one of the most important data analysis methods providing a platform for advanced chemometrics methods. As stated in (S. Wold *et al*, 1987) PCA can have many goals; simplification, data reduction, modelling, outlier detection, variable selection, classification, prediction and unmixing. It can be used to understand general characteristics of data set and guide further investigation through more refined techniques (Wentzell *et al*, 2012). PCA reduces the dimension of data by calculating principal components (PCs) which reflect the structure of data corresponding to maximum variance. These PCs can then be plotted to visualize the relationship between samples and variables through the use of scores and loading plots. A tutorial review on PCA can be found in (Bro *et al*, 2014). Decomposing a data matrix X into a structure part which consists of a score matrix (T) and a

loading matrix (P) and noise part or residual matrix (E), is shown in (4) and (5).

$$X = TP^T + E \quad (4)$$

$$X = t_1 P_1{}^T + t_2 P_2{}^T + t_3 P_3{}^T + \cdots + t_A P_A{}^T + E \quad (5)$$

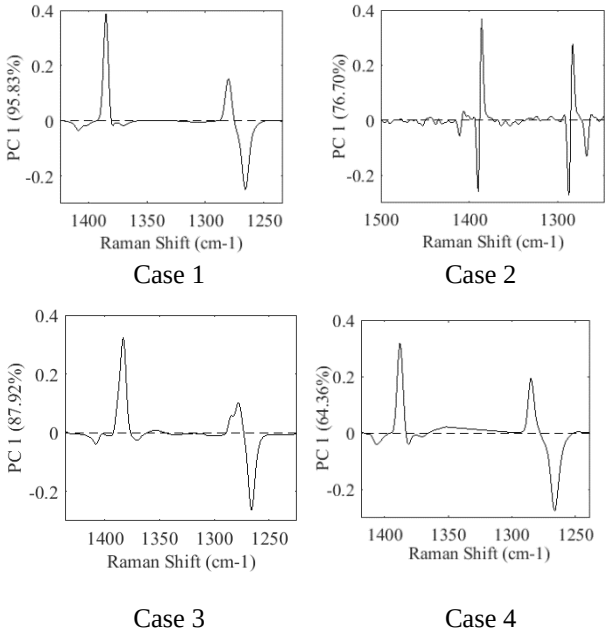


Figure 5. Loading plots of the first principal component for case 1-4 (Region : 1225 – 1450 cm-1).

t_A and P_A are score vector and loading vector for PCA respectively. PC1 is the first principal component which relates to the maximum variance of the data, and PC2 is the second principal component which corresponds to the second largest variance etc. Score values provide information about sample variations while loading value explains the relationship between variables. Residuals provides information as to what spectral variations have not been explained. There are different ways to decompose a matrix to score and loading vectors. NIPALS (*Non-linear Iterative Partial Least Squares*) algorithm (H. Wold, 1966) uses iterative sequence of ordinary least square regression to calculate PCs and was used in this study.

2.6 Important variables related to CO_2 absorption

CO_2 absorbed PO, PPC and amine mixtures exhibit several sharp overlapping peaks in the region of 300 to 1500 cm-1 and 2700 to 3250 cm-1. The focal point in this study is to investigate CO_2 absorption and hence only the peaks related to absorbed CO_2 are considered in the model development. In case 3-4, the monomer PO and the polymer PPC were added into the autoclave reactor and the CO_2 was absorbed into this mixture. Therefore, CO2 bands related to dissolved CO_2 in the PO or PO/PPC mixture were followed in this study.

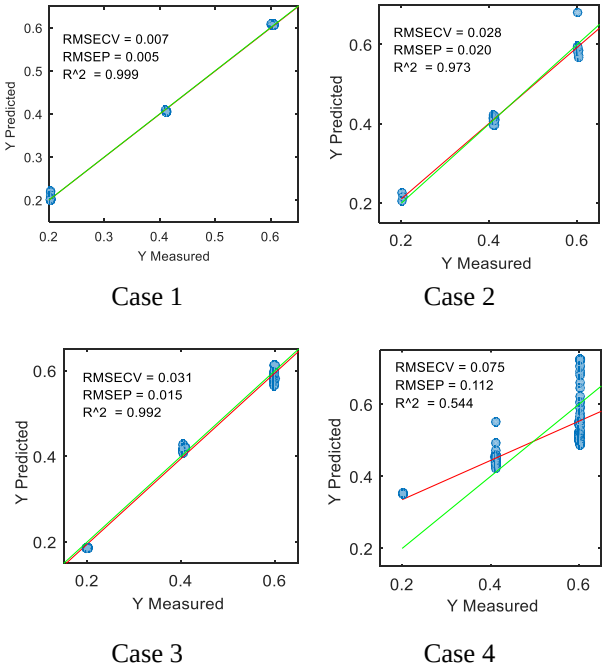


Figure 6. Development of linear regression model using PC1 score values and thermodynamic model data.

(Y measured = CO_2 mole fraction predicted by VLE data; Y predicted = CO_2 mole fraction predicted by PC1 scores; red line= best fitted line based on calibration points; green line=1:1 target line; RMSE (CV/P)= root mean square error of (cross validation/prediction)

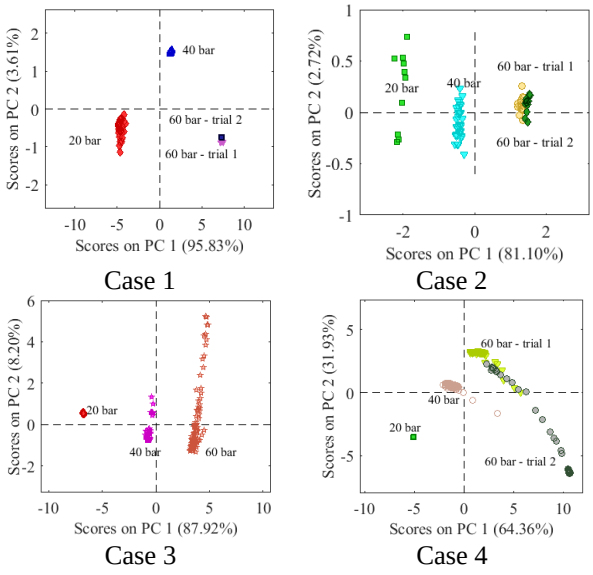


Figure 7. Score plots – PC1 vs PC2 for case 1-4.

Literature reports such Raman wavenumbers of 1264, 1284, 1369, 1387, 1408 cm-1 (Hanf *et al*, 2014). In case 5-8, peaks related to carbonate, bicarbonate, carbamate and dissolved CO_2 fall in the region of 1000-1500 cm-1 ((Vogt *et al*, 2011), (Wong *et al*, 2015)). Therefore, for development of PCA models, the region between 1000-1500 cm-1 and 1225-1450 cm^{-1} were selected for case 1-4 and case 5-8 respectively.

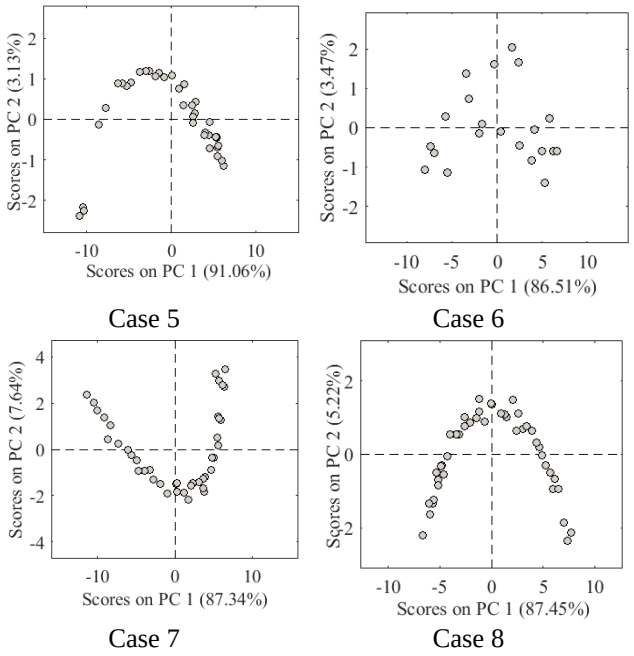


Figure 8. Score plots for case 5-8.

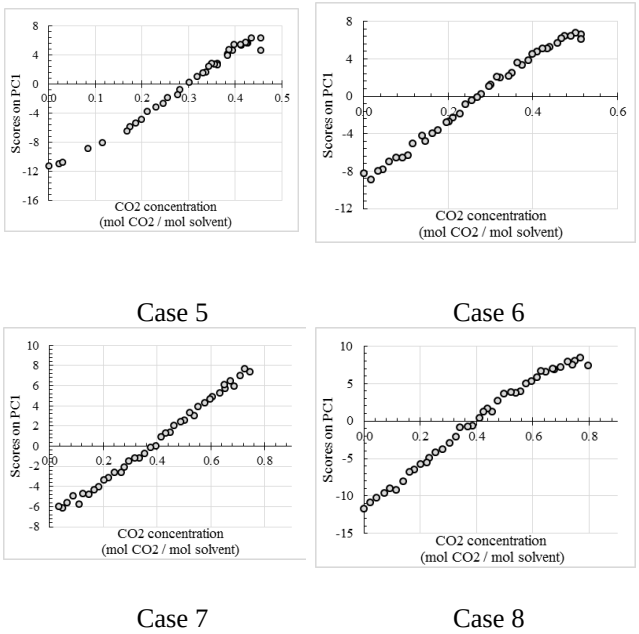


Figure 10. Development of linear regression model using PC1 score values and CO_2 content given by titration data.

3 Results and Discussion

Figure 2 (a) and (b) show raw spectra for case 1 and 3 respectively highlighting spectral variation with increasing pressure in the region of 1225 to 1450 cm^{-1}. CO_2 peaks at 1264, 1284, 1369, 1387, 1408 cm-1 can be identified in this figure. A similar spectral behavior was observed for case 2 and 4 in the same region. Figure 3 gives raw Raman signals observed for CO_2 loaded 4 different amine solvents. Only two spectra from each solvent are shown. Figure 4 shows how the intensity of Raman bands varies with the CO_2 content for MEA samples (Case 5). Both Figure 3and Figure 4 claim that spectral evolution in the region between 1000 to 1500 cm-1, for case 5-8 with respect to case 1-4 is complex due to curved baseline, baseline offsets and overlapping bands. The reason is that the chemical products when CO_2 is reacted with the solvent appear with overlapping peaks in this region. Therefore, when quantifying the total amount of CO_2 absorbed in solvent, all these peaks are needed.

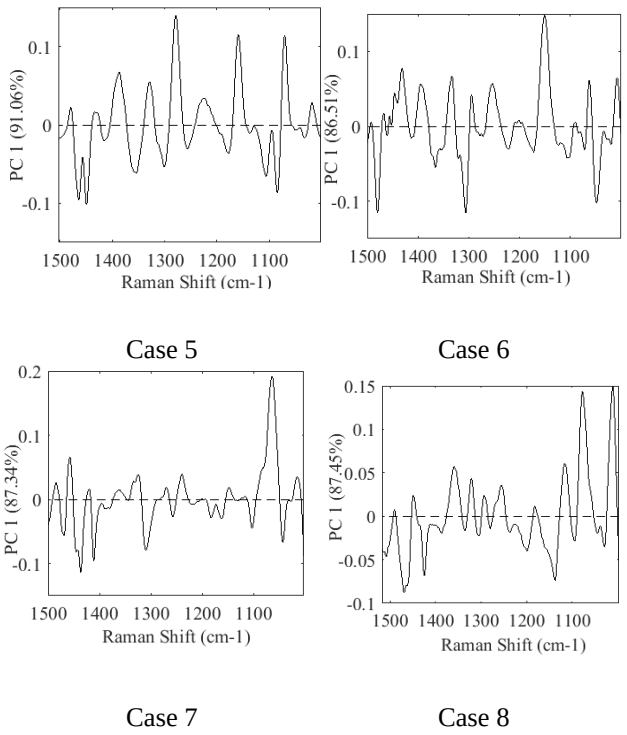


Figure 9. Loading plots of the first principal component for case 5-8.

All the Raman signals under each case were first smoothed using baseline-whittaker smoother, then SNV and finally mean centered. PCA was performed for processed data. First principal component was identified as the dimension explaining the largest variance of data in each case. Finally, score values of PC1, were compared with the mole fraction of CO_2 predicted by thermodynamic models for case 1-4 and CO_2 amount determined from titration for case 5-8. Loading plot, score plot and comparison of PC1 score value with CO_2 content under each case were used to explain characteristics in each system.

3.1 Case 1-4

With reference to Figure 5 loading plots of case 1, 3 and 4 almost give similar information about important variables (Raman shifts) while case 2 is different. This is caused by exposing the Raman sensor to both gas and liquid phases as a result of high stirrer speed and

development of vortex in case 2. There is also low viscosity in the medium at low pressures, which creates high turbulence. Score plots of PC1 vs. PC2 as given in Figure 7, show clear distinguish of recorded signals between the three pressure values of 20, 40 and 60 bar. PC2 direction explains only a small variation of data for case 1-3. Experiments for 60 bar, were conducted in replicates and their overlap in score values could be observed in case 1 and 2.

Figure 6 shows how closely PC1 score values are related to VLE data. Plots in this figure were derived by linear regression between PC1 score values as X variables and predicted CO_2 content from VLE data as y variables. From VLE data, CO_2 mole fractions at 20, 40 and 40 bar are 0.202, 0.411 and 0.601 respectively. These values are represented as 'Y measured' in Figure 6. PC1 score values at 3 pressure conditions follow the linear trend given by the mole fraction of CO_2 predicted by thermodynamic models at case 1 and 2. In the presence of PPC (case 3-4), even though pressure and temperature were maintained constant, a significant time was needed to achieve equilibrium condition of CO_2 absorption by the solvent especially at higher pressure region. For example, at 40 bar and 60 bar, PC1 score value of the initial spectra is less than the final recorded spectra at that condition. Therefore, although the reactor is maintained at the required pressure, the score plot gives the hint whether the equilibrium condition has been achieved or not. The significance of the above fact can be clearly understood when examining the score plot for case 4 (Figure 7). In this trial, we see that only 20 bar condition shows a compressed data swarm while at 40 bar, PC1 score values increases with time and this variation is more significant for 60 bar. This is further assured by Figure 6 (case 4) where the thermodynamic model satisfies the trend of final recorded data for 60 bar condition, but highly deviate from the initial recorded data at this condition. PC1 score values positively correlate with the amount of absorbed CO_2 by PO-CO_2 and PO-PPC-CO_2 systems.

3.2 Case 5-8

Absorption of CO_2 by amines (case 5-8), features several important variables in the region 1000-1500 cm-1 as given by loading plots in Figure 9 and this is the result of several parallel equilibrium reactions happening in each system. Each sample carries different information which mean different amount of CO_2 absorption and hence the concentration of chemical species produced during these reactions are different. Therefore, a data spread in score plot of PC1 vs PC2 can be observed in the score plots in each case as presented in Figure 8. However, similar to polymer-CO2 system, PC1 explains the largest variation of data structure and therefore PC1 score values were compared with total CO_2 absorbed by the system. Results are shown in Figure 10. With the increasing amount of CO_2, there is a gradual increase of PC1 score value highlighting that PC1 score value is an indication of the level of CO_2 absorbed by the sample.

4 Conclusions

Monitoring CO_2 in liquid phase of PO-CO_2 system or PO-PPC-CO_2 system by analytical techniques is challenging as the CO_2 quickly desorbs if the pressure is lowered in sample taking. Therefore, online analysis such as spectroscopy is more favorable For CO_2-amine systems, an in situ characterization of CO_2 absorption gives credits to process monitoring and control ability. Based on this study, combination of Raman spectroscopy and PCA claims that PC1 score value explains variation of data structure corresponded to absorbed CO_2 amount. PCA plots give an indication of CO_2 composition, process dynamics and equilibrium conditions in these two chemical systems and hence can be used as an efficient tool to analyse collinear process data. Further investigation of PCA model development under different process parameters is recommended to validate the findings from this feasibility study. Experiments to develop advanced chemometrics tools such as partial least square regression can now be recommended for both polymer-CO_2 system and amine-CO_2 system.

References

C. Atzberger and P. H. C. Eilers. Evaluating the effectiveness of smoothing algorithms in the absence of ground reference measurements *International Journal of Remote Sensing*, 32(13): 3689–3709, 2010. doi:10.1080/01431161003762405

R. J. Barnes, M. S. Dhanoa, and S. J. Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43(5): 772-777, 1989. doi:10.1366/0003702894202201

R. Bro and A. K. Smilde. Centering and scaling in component analysis. *Journal of Chemometrics*, 17(1): 16–33, 2013. doi:10.1002/cem.773

R. Bro and A. K. Smilde. Principal component analysis. *Anal. Methods*, 6: 2812-2831, 2014. doi:10.1039/C3AY41907J

L. B. Chen and X. G. Fang. Phase equilibria and CO2 distribution in the CO2-epoxide-toluene systems. *Natural Gas Chemical Industry*, 3: 255-266, 1994.

P. H. C. Eilers. A Perfect smoother. *Analytical Chemistry*, 75(14): 3631-3636, 2003. doi:10.1021/ac034173t

S. Hanf, R. Keiner, D. Yan, J. Popp, and T. Frosch. Fiber-enhanced Raman multigas spectroscopy: A versatile tool for environmental gas sensing and breath analysis. *Analytical Chemistry*, 86(11): 5278-5285, 2014. doi:10.1021/ac404162w

X. Jiang, F. Gou, and H. Jing. Alternating copolymerization of CO2 and propylene oxide catalyzed by C2v-porphyrin cobalt: Selectivity control and a kinetic study. *Journal of Catalysis*, 313: 159-167, 2014. doi:10.1016/j.jcat.2014.03.008

D. Y. C. Leung, G. Caramanna, and M. M. Maroto-Valer. An overview of current status of carbon dioxide capture

and storage technologies. *Renewable and Sustainable Energy Reviews*, 39: 426-443, 2014. doi:10.1016/j.rser.2014.07.093

N. McCann, D. Phan, X. Wang, W. Conway, R. Burns, M. Attalla, G. Puxty, and M. Maeder. Kinetics and mechanism of carbamate formation from CO2(aq), carbonate species, and monoethanolamine in aqueous solution. *The Journal of Physical Chemistry A*, 113(17): 5022–5029, 2009. doi:10.1021/jp810564z

S. F. Shakhova, O. L. Rutenberg, and M. N. Targanskaya. Liquid-gas equilibrium in the epoxypropane–carbon dioxide system. *Russ. J. Phys. Chem*, 47(6), 1973.

M. Vogt, C. Pasel, and D. Bathen. Characterisation of CO2 absorption in various solvents for PCC applications by Raman spectroscopy. *Energy Procedia*, 4: 1520-1525, 2011. doi:10.1016/j.egypro.2011.02.020

P. D. Wentzell and S. Hou. Exploratory data analysis with noisy measurements. *Journal of Chemometrics*, 26(6): 264–281, 2012. doi:10.1002/cem.2428

H. Wold. Estimation of principal components and related models by iterative least squares. *Journal of Multivariate Analysis* 1: 391-420, 1966.

S. Wold, K. H. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1): 37-52, 1987. doi:10.1016/0169-7439(87)80084-9

M. K. Wong, M. A. Bustam, and A. M. Shariff. Chemical speciation of CO2 absorption in aqueous monoethanolamine investigated by in situ Raman spectroscopy. *International Journal of Greenhouse Gas Control*, 39: 139–147, 2015. doi:10.1016/j.ijggc.2015.05.016

Modeling and Portfolio Optimization of Stochastic Discrete-Event System through Markovian Approximation: an Open-Pit Mine Study

Roberto G. Ribeiro[1] Rodney R. Saldanha[2] Carlos A. Maia[2]

[1]Graduate Program in Electrical Engineering, Federal University of Minas Gerais - UFMG, Brazil,
rogorib@gmail.com
[2] Dep. of Electrical Engineering, Federal University of Minas Gerais - UFMG, Brazil,
{rodney,maia}@cpdee.ufmg.br

Abstract

Operation in an open-pit mining is a complex task with stochastic nature. Usually, this kind of system is analyzed by means of DES (Discrete Event System) simulation. This work considers optimizing the investment in new projects in such a way to reach maximum production of an open-pit mine. When a DES model is associated with an optimization problem, the time taken to run such model is a crucial aspect. In order to analyze the project impacts in a reasonable time, this work presents a DES markovian model which represents a load-haulage cycle. The results obtained were compared with the results acquired from validated simulation models which represent the same system. In the optimization context, the complexity is exponential. Therefore, this work proposes a formulation that considers the inter-relationship between projects, which aims to help decision makers. Instead of trying all the possible projects combinations, the proposed method searches for identifying the set of projects that produce good feasible solutions based on performance measure from designed DES model.

Keywords: Markov chain, closed queuing network, project portfolio, open-pit mining.

1 Introduction

Typically, Project Portfolio Optimization decisions are strategic decisions that are made on a yearly basis. For open-pit mine, investments in its operations are needed to reduce the production cost, enhance production, eliminate bottlenecks and to improve the use of system resources. Thus, in a Brazilian mining, several projects are proposed annually in order to improve the company competitiveness. The Project Portfolio consists of many actions, including road improvements, truck or crane acquisition, hire operator, etc. Each project impacts in a specific indicator of the open-pit mine. A preliminary goal is to identify how these projects influence the company competitiveness. A fundamental goal of any mining project is maximizing the total mine ore production in a specific time horizon. Nevertheless, to improve some indicator cannot result in an effective contribution for this goal.

The open-pit-mine is a complex and stochastic system, in which interactions between several agents impact heavily on the total mine ore production. Usually, this kind of system is analyzed by means of DES simulation. According to (Banks, 2000), simulation is the imitation of the operation of a real-word process or system over time, involving the generation of an artificial history of the system. In order to analyze the real impact of each project, a DES simulation model can be used. Therefore, it is possible to estimate and make better decisions.

When DES simulation model is associated with a discrete decision problem, we can associate the control variables to a discrete optimization problem. Discrete optimization with simulation is a methodology known in the literature as *discrete optimization via simulation*. According to (Nelson, 2010), this methodology addresses solving problems with a countable number of feasible solutions, when the system is complex enough that its expected value is estimated by running simulation.

Most commercial DES simulation software are associated with some optimization software. However, these softwares use meta-heuristics or heuristics that, although efficient in many cases and modeled to be generic and applicable in various contexts, it does not exploit the problem features and hence tend to be less efficient. Fu (Fu, 2002) summarizes the optimization packages used by most popular commercial DES simulation softwares and the search strategy used as well. According to (Fu, 2002), algorithms that apply a very general way often have a slow convergence in practice.

In a portfolio optimization, the goal is to determine a set of projects that maximizes some indicator, such as *total mine production* index. Each portfolio is formed at least by one project and correct projects combinations can maximize the expected return. From the optimization point of view, the number of combinations has an exponential growth. Thus, usually it is not possible analyze all the possibilities, especially when each evaluation is obtained using DES simulation, since this methodology can be very burden.

According to (Bolch et al., 1998), the main drawback of DES simulation models is the time taken to run such models for large and realistic systems particularly when results with high accuracy are desired. Thereby, combining DES simulation with optimization is a major challenge.

Taken into account the difficulties inherent to discrete optimization via simulation, two approaches to research can be explored. The first consist in reducing the computational time used in the function evaluation. When we consider a DES simulation model as an objective function of an optimization problem, the time taken to run such model is a crucial aspect. The second approach consist in developing a specific optimization method to the problem that explores peculiarities of the modeled system.

1.1 First step - DES model

The first step aims to create a modeling of stochastic DES through markovian approximation. This strategy was motivated by successful application in the field of the performance evaluation of computational system. According to (Bolch et al., 1998), a cost-effective alternative to DES simulation consists of analytic models, which can provide relatively quick answers and more insight to the system being studied. Before the creation of nowadays computational technology, problems with stochastic nature were generally solved analytically. Marie (Marie, 2011) claims that scientific ambition was limited by computing power, i.e., it was necessary to use the imagination to look for approximations in order to reduce a state space to a few hundred states. Recently, (Marie, 2011) observed that huge amounts of available computing resources increase the trend to solve models through simulation and did not encouraged researchers to look for tractable analytical solutions.

It is important to point out that simulation is enshrined as a good methodology to treat DES model. However, in the optimization context, the number of evaluations depends on the time taken by run the designed model. Reference (Ekren et al., 2013) presents an analytical model for an autonomous vehicle storage and retrieval system. The authors model a material-handling system as a semi-open queuing network to be used instead of DES simulation. According to (Ekren et al., 2013), the analytical model is useful in estimating key performance measures of alternate configurations of the system quickly and accurately.

Marie (Marie, 2011) explain that some mathematical/probabilistic properties can be used to analyze problems of stochastic nature without simulation. Therefore, the first step of this study consist to explore and to apply this properties. A load-haulage cycle of a realistic Brazilian open-pit mine is considered. This cycle is modeled as a closed queuing network where all queues (also known as nodes) are connected. The idea is to measure the *mean response time* of each node, and consequently, the *mean cycle time* and the *total mine production* index. Therefore, these performance measures are used in the optimization context to find a good and fast answer.

In this study we consider the cdf (Cumulative Distribution Function) approximation using exponential distribution. We assume that the service time of each node is exponentially distributed while this is not true in the real model. It is important to point out that this kind of approximation cannot provide a reasonable accuracy. Therefore, the results obtained by this approach are compared with the results acquired using validated simulation models presented in (Ribeiro, 2015). These models consider pdf (Probability Density Function) approximation using general distribution. A Petri net model and a SIMAN model representing the same load-haulage cycle are considered. The assumption is to verify whether there are significant difference among the designed analytical approximation model and the mentioned methodologies of DES simulation.

1.2 Second Step - Portfolio optimization

The second step of this study consist in determining the set of projects which maximizes the *total mine production* index respecting the established budget. One approach is to formulate this decision as a knapsack problem which is a classical combinatorial optimization problem (Pisinger, 1994). Figuratively, we can describe this problem as filling a backpack without exceeding a certain volume limit. The decision consists to place in the backpack products that maximize (maximization problem) a specified value, respecting its capacity. In this study, the products are the projects at the portfolio and the capacity is the available budget.

In order to analyze the real impact of witch project, the DES model is used. Hence, it is possible to estimate how much each one increases the *total mine production* index. Here, we named this estimative as 'gain'. A linear solution can be found maximizing the sum of individual gain. However, we cannot consider each project individually. Due to the open-pit mine features, there are interrelationships among projects. It means that a combination of two or more projects must result in different gain compared with the linear solution. Since it may unfeasible evaluate all projects combinations, the optimization strategy aims to create a formulation where the decision variables set is formed by all individuals projects in the portfolio and all 'relevant combinations'. In this study, a 'relevant combination' is a set with at least two projects with a strong inter-relationship among them. A major challenge is to find how intense this inter-relationship is.

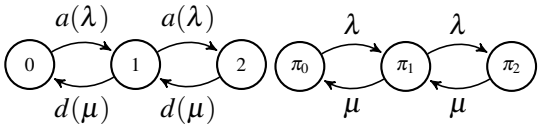
In the load-haulage cycle, the *total mine production* index depends on the *mean response time* taken in each node of the cycle. Some projects aims to reduce the service time in a specific node. A disadvantage of applying these projects is the possibility of increasing the *mean response time* of others nodes. Consequently, neither total cycle time nor *total mine production* index will be changed. The central core of the optimization method is to use the *mean response time* of each node to find 'relevant combinations'. The major idea is that this performance measure

indicates the plant bottlenecks, which must be fixed by the new projects. Section 5 describes the philosophy behind of the strategy.

2 Stochastic DES model through markovian properties

A general way to represent a stochastic DES model is through stochastic timed automata. According to (Cassandras, 2008), any timed model for DES requires the specification of a clock structure. This structure is a set of distribution functions, one for each event, characterizing the corresponding lifetimes. Beside that, it determines the dynamic of the system. When all clock sequences are iid (Independent and Identically Distributed) and the interevent times are exponentially distributed, a stochastic DES can be modeled as a Markov Chain (Cassandras, 2008).

Figure 1a shows a stochastic timed automata where a and d are events with occurrence times rates of λ and μ, respectively. Considering the markovian properties presented in (Cassandras, 2008), this automata generates a Markov chain illustrated in Figure 1b.



(a) Stoch. timed automata **(b)** Markov chain

Figure 1. Stochastic timed automata as Markov Chain

In Figure 1b, π_0, π_1 and π_2 denotes the stationary probability of the states 0, 1 and 2, respectively. According to (Cassandras, 2008), a system modeled as a Markov chain is allowed to operate for a sufficiently long period of time so that the state probabilities can reach some fixed values which no longer vary with time. The main objective of using a Markov chain is to compute these probabilities. According to (Bolch et al., 1998), long run dynamics of Markov chains can be studied using a system of linear equations with one equation for each state. Thus, solution of these equations results in stationary probabilities of the Markov chain, consequently, desired performance measures such as *mean response time* can be easily obtained.

It is known that a DES simulation model can be seen as a parallel composition of many stochastic timed automata. The set of clock structures describes how the DES model evolves as a result of event occurrences over time. Considering that all clock structure of this DES model conforms to the markovian properties, the full DES model can be simplified to a Markov chain (In a parallel composition model, the stationary probabilities are denoted by the M-tuple $\pi(n_0,\ldots,n_M)$ where M represents the number of individual automata). Accordingly, a system of linear equations to compute the stationary probabilities must still be valid. The problem of this representation is the fact that the cardinality of the state space can grow drastically in a Markov chain of a complex system. Therefore, computing the stationary probabilities of the states became a hard (even impossible) task.

Fortunately, for a class of Markov Chain, which can be expressed as a of queuing network, very fast numerical solution methods have been developed to derive important performance measures without resorting to the underlying state space (Bolch et al., 1998). One of these methods is known as convolution algorithm. This method is a sufficient analytical technique to obtain the interest measures of this study. This algorithm aims to evaluate an important performance measure without having to compute explicitly the stationary probabilities.

2.1 Product-Form Networks - Convolution algorithm

Once we approximate the stochastic automata by a Markov Chain and then by a queuing network, to compute the stationary probability it is necessary to consider the theorem proposed by (Jackson, 1957). The Jackson's theorem (also known as product-form) says the steady-state probability of the network can be expressed as the product of the state probabilities of the individual nodes. Based in the Jackson's theorem, (Gordon and Newell, 1967) present a general formulation to compute the stationary probabilities in a closed queuing network. This formulation follows the Eq. 1:

$$\pi(n_1,\ldots,n_M) = \frac{1}{G(N)} \prod_{m=1}^{M} \beta_m(n_m), \tag{1}$$

where $G(N)$ denotes the normalization constant. In this equation, $\beta_m(n_m)$ denotes a step function which depends on the number of customer n_m in each node m. Let μ_m be the service rate of the node m, and let ν_m be the relative arrival rate of the same node, the value of $\beta_m(n_m)$ is obtained using Eq. 2.

$$\beta_m(n_m) = \begin{cases} 1, & \text{if } n_m = 0 \\ \dfrac{\left(\frac{\nu_m}{\mu_m}\right)^{n_m}}{\prod_{k=1}^{n_m} b_m(k)}, & \text{else} \end{cases} \tag{2}$$

Usually, in a closed queuing network μ_m is known, while ν_m should be compute. Thus, the balance equations Eq. 3 can be used to compute ν_m, where the parameter p_{im} denotes the transition probability that a customer departs from node i and arrives at the node m.

$$\nu_m = \sum_{i=1}^{M} \nu_j p_{im} \quad \forall\, m = 1,\ldots,M. \tag{3}$$

The step function Eq. 2 depends on the queuing discipline of the node as well. In this study we consider the disciplines: FIFO (First In First Out) and IS (Infinity Server). Thus, $b_m(k)$ denotes a step function which depends on the

queuing discipline and the number of service q_m, as we can see in Eq. 4.

$$b_m(k) = \begin{cases} 0, & \text{if } k = 0 \\ \min(k, q_m), & \text{if } 0 < k \leq N \ \ \& \ \ \text{FIFO} \\ k, & \text{if } 0 < k \leq N \ \ \& \ \ \text{IS} \end{cases} \quad (4)$$

According to Buzen algorithm (Buzen, 1973), the normalization constant $G(N)$ can be computed following Eq. 5.

$$g(n,m) = \begin{cases} 1, & \text{if } n = 0, \ \forall \ m \\ \beta_m(n), & \text{if } m = 1, \ \forall \ n \\ \sum_{k=0}^{n} \beta_m(k) g(n-k, m-1), & \text{else} \end{cases} \quad (5)$$

In the Eq. 5, $g(n,m)$ denotes the normalization constant of all possible sub closed queuing network where n ($0 \leq n \leq N$) and m ($1 \leq m \leq M$) indicate, respectively, the amount of clients and nodes. Therefore, $g(N,M)$ represents $G(N)$.

2.2 Marginal probability

In order to compute the *mean response time*, it is necessary to evaluate the marginal probability. This performance measure $\pi_m(n)$ represents the probability of having n clients in the node m, accordingly with Eq. 6.

$$\pi_m(n) = \sum_{\substack{\pi_{n_1} \ldots \pi_{n_M} \\ \& \ n_m = n}} \pi(n_1, \ldots, n_M) \quad (6)$$

As mentioned, the number of states in a Markov chain of a complex system can become very huge. Therefore, it cannot be possible to compute all $\pi(n_1, \ldots, n_M)$. However, the convolution algorithm described in (Bolch et al., 1998) is sufficient to obtain the marginal probability directly. In general terms, this technique consist of to substitute Eq. 1 in Eq. 6, which results in Eq. 7.

$$\pi_m(n) = \frac{\beta_m(n)}{G(N)} G_M^{(m)}(N-n) \quad (7)$$

The constant $G_M^{(m)}(N-n)$ must be interpreted as normalization constant of a closed queuing network without the node m and with n clients less. The value of $G_M^{(m)}(N-n)$ can be computed recursively (Bolch et al., 1998), following Eq. 8.

$$G_M^{(m)}(k) = \begin{cases} 1, & \text{if } k = 0 \\ G(N) - \sum_{k=1}^{N} \beta_m(n) G_M^{(m)}(N-k), & \text{else} \end{cases} \quad (8)$$

2.3 Mean response time

As mentioned in section 1, the idea is to measure the *mean response time* of each node. Following the Little's law, this performance measure is obtained using Eq. 9:

$$\bar{T}_m = \frac{\bar{N}_m}{\psi_m}, \quad (9)$$

where $\bar{N}_m$ and ψ_m are the *mean number of clients* and the *throughput* of the node m, respectively. Moreover, these performance measures must be computed by Eq. 10 and Eq. 11, where q_m denotes the number of services in the node m.

$$\bar{N}_m = \sum_{n=1}^{\infty} n \pi_m(n) \ \ \forall \ 1 \leq m \leq M \quad (10)$$

$$\psi_m = \sum_{n=1}^{\infty} \pi_m(n) \min(n, q_m) \mu_m \quad (11)$$

In this section, we show an efficient technique to compute the *mean response time*. This performance measure can be obtained analytically in a quiet fast. In the following section we introduce concepts about the load-haulage cycle and demonstrates how to use the *mean response time* to compute the *total mine production* index.

3 Load-haulage cycle

In (Torkamani and Askari-Nasab, 2012), a DES simulation model was implemented in SIMAN language (Pegden, 1983) in order to simulate the load-haulage system in an open-pit mine. Each simulation scenario used a distinct combination of the number of trucks and shovels. The goal was maximizing the mine production index at lowest possible operating cost. As the use of simulation requires high computational effort, sometimes it is not possible to try all feasible scenarios. Thus, the strategy taken by (Torkamani and Askari-Nasab, 2012) considers proper indicators to choose the scenarios to be evaluated.

In (Ercelebi and Bascetin, 2009), another strategy for allocation of trucks in open-pit mine was proposed. The authors applied closed queuing theory to obtain some measures of interest such as the *total mine production* index. According to (Ercelebi and Bascetin, 2009), in a load-haulage system, the *total mine production* index P over a given time period of interest can be estimated by Eq. 12:

$$P = N \cdot C \cdot \frac{T_{horizon}}{\bar{T}_{cycle}}, \quad (12)$$

where C denotes the truck's capacity, N the number of trucks and $T_{horizon}$ the period of interest. The measure $\bar{T}_{Cycle}$ represents the *mean cycle time*.

In a load-haulage system, basically, each truck goes to load site and waits until the loading process is completed. Following, the trucks go to the dump site and dump the ore into a crusher. However, each truck can be diverted to other process during the transport, such as maintenance, supply, etc. Known as operational stops, these processes must be included in the load-haulage cycle with a transient probability associated p_m. The Eq. 13 represents a general expression to compute the *mean cycle time* considering the operational stops:

$$\bar{T}_{cycle} = \sum_{m=1}^{M'} \bar{T}_m + \sum_{m=M'+1}^{M} p_m \bar{T}_m, \quad (13)$$

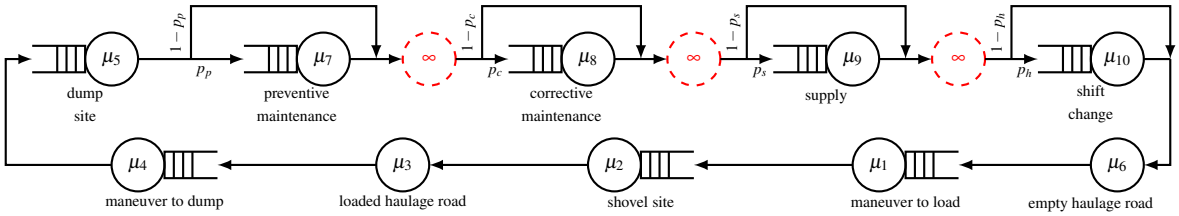


Figure 2. Closed queuing network representing the Load-haulage system

where the label m, from 1 to M' denotes the basic process of the load-haulage system, while m from 1 to M' indicates process which are operational stops.

4 DES model: Load-haulage system

In order to compute the *total mine production* index, the analytical technique presented in this paper was tested in a load-haulage system depicting a mining front of a brazilian open-pit mine. As mentioned in section 1, in this study we consider the cdf approximation using exponential distribution. The assumption is verifying whether there are significant difference among the results obtained and the results presented in (Ribeiro, 2015). Table 1 presents the process of the load-haulage system. The pdf, the first moment approximation and discipline are showed as well.

Table 1. First moment approximation

Process	Dist.	$E[x]$	μ_m	Dis
Maneuver to load	Triangular	2.266	$\frac{1}{2.266}$	FIFO
Shovel site	Tria (comp.)	3.227	$\frac{1}{3.227}$	FIFO
Loaded Haulage road	Inv. Gaussian	8.333	$\frac{1}{8.333}$	IS
Maneuver to Dump	Triangular.	1.100	$\frac{1}{1.100}$	FIFO
Dump site	Triangular	2.133	$\frac{1}{2.133}$	FIFO
Empty Haulage road	Inv. Gaussian	6.944	$\frac{1}{6.944}$	IS
Preventive maintenance	Triangular	150	$\frac{1}{150}$	FIFO
Corrective maintenance	Triangular	1120	$\frac{1}{1120}$	FIFO
Supply	Gaussian	150	$\frac{1}{150}$	FIFO
Shift Change	Gaussian	16	$\frac{1}{16}$	FIFO

The load-haulage system depicting a mining front was modeled as a queuing network. Thus, each process presented in Table 1 is seen as a node and each truck is a client. Figure 2 shows the full model. As we can see, there are four operational stops in this model. The transient probabilities to these nodes are: p_p, p_c, p_s and p_h. In addiction, it was added to the model three fictitious nodes (dashed nodes) to establish transient decisions between operational stop nodes. Taken to account that these nodes have infinity rate, they do not change the final result.

The transient probabilities were computed using the occurrence period of each operational stop of the open-pit mining front. Therefore, the estimated values are: $p_p = 0.0015$, $p_c = 0.0127$, $p_s = 0.0255$ and $p_h = 0.1232$.

Considering the transient probabilities values, the service rates and the queuing discipline showed in Table 1 the *mean response time* of each node is obtained using the technique presented in section 2. Consequently, using Eq. 13 we have the Eq. 14.

$$T_{cycle} = \sum_{m=1}^{6} \bar{T}_m + p_p\bar{T}_7 + p_c\bar{T}_8 + p_a\bar{T}_9 + p_t\bar{T}_{10} \qquad (14)$$

As mentioned before, the aim is to compute the *total mine production* index. However, it necessary to estimate the parameter C (truck's capacity). In this study, it was considered a caterpillar 793F truck with nominal payload capacity of 226.8 tonnes. Finally, specifying a time period of interest, the mine production index must be evaluated by Eq. 12.

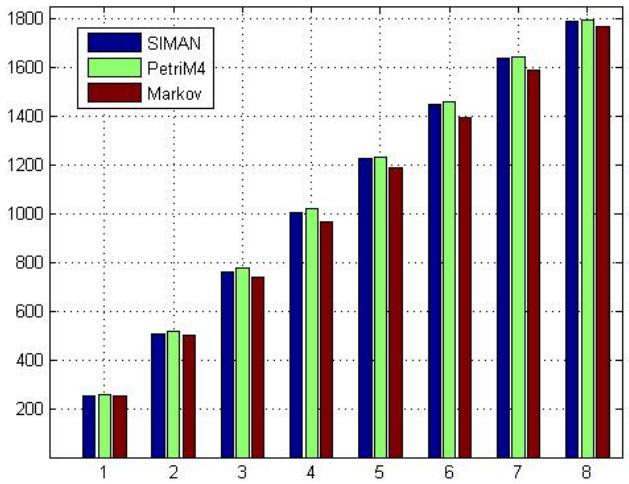


Figure 3. N^o of trucks vs Mine production index -tonnes(mil)

For a time period of interest of one month, the closed queuing network was evaluated changing the number of trucks N (clients), from 1 to 8. Figure 3 shows the comparison among the results obtained and the results presented in reference (Ribeiro, 2015).

From Figure 3 it is possible to observe there is no significant difference among the analytical approximation model designed and the methodologies of DES simulation, neither SIMAN nor Petri Nets. The maximum difference between the analytical approximation model and the

SIMAN model was 4%, while the maximum difference when compared with the Petri net model was 5.16%.

It is important to emphasize that the analytical approximation model runs hundreds times faster than DES simulation models. In conclusion, the DES through markovian properties consist of a reasonable way to represent the load-haulage system in study.

5 Project portfolio formulation

Let P_0 be the *total mine production* index of the load-haulage system shown in Figure 2, with 8 caterpillar 793F trucks and without any project application. Using Eq. 12 we have $P_0 = 1767.29$ tonnes. Let P_k be the *total mine production* index of the same system with the application of the project k, and let $E[g_k]$ be the expected gain provided by the application of this project, that is $E[g_k] = P_k - P_0$.

Knapsack problem is a typical formulation for portfolio optimization. In this study, the goal consists of maximize the sum of the gains, respecting the established budget. However, it is necessary to consider the inter-relationship between two or more projects. Since it may be unfeasible to evaluated all project combinations, the *mean response time* of each node is useful to define those who deserve to be evaluated. The major idea is that this performance measure indicates the plant bottlenecks, which must be fixed by the new projects.

Initially, consider that two or more projects have a small expected gain. Using conventional optimization methods they would hardly be included in the final optimal portfolio. However, the combination of them can generate good expected gain when evaluated jointly. Considering these circumstances, the analysis of the problem is necessary to find which combinations are relevant.

For example, a project that produces a time reduction in the dumping process and increases *mean response time* in the shovel site can produce a significant gain whether combined with another projects that provides an improvement in the second node.

A relevant combination could be composed by more than two projects. In this methodology each combination just can be obtained by pairs. However, combinations found and evaluated are converted in new decision variables (new projects). Then, in the next iteration, the strategy can combine this new project with another one.

The Eq. 15 presents a formulation where the set of decision depends of the number of projects combination $\mathbf{S} = \{X_0, \ldots, X_k, \ldots, X_K\}$. Since X_k denotes the *k-th* combination evaluated, x_k is a decision variable that represents it in the optimization function.

Considering that in equation 15 the optimal solution can be composed by more than one project combination, it is necessary to prevent that a isolated project appear redundantly in the same solution. For example, suppose we have the combinations: $X_1 = \{p_1, p_2\}$ and $X_2 = \{p_2, p_3\}$. In this circumstances, it is necessary to append at the formulation (Eq. 15) the constraint $x_1 + x_2 \leq 1$. In another case, assume that we have two other project combination previously evaluated: $X_3 = \{p_3, p_4\}$ and $X_4 = \{p_1, p_2, p_3, p_4\}$. Consequently, it is necessary to add the constraint $x_1 + x_3 \leq 1$ because a combination between this two set is equal to X_4. Based in this rules, the formulation must account the preposition 1:

Table 2. Project portfolio

Proj	Description	c (pu)	Impact	Where?
p_1	Bilateral charging	0.5	-15%	node 1
p_2	Roads improvement	1.5	-20%	nodes 3,6
p_3	Dead load reduction	1	5%	par. C
p_4	Excavation	2	-18%	node 2
p_5	1 Truck acquisition	5	$N+1$	par. N
p_6	2 Truck acquisition	10	$N+2$	par. N
p_7	Rolling A-Frame	1	-17%	node 8
p_8	Forklif acquisition	0.7	-15%	node 8
p_9	Preventive kits	0.65	-14%	node 8
p_{10}	Supply improvement	2	-35%	node 9
p_{11}	Dump site improvement	2	-25%	node 4
p_{12}	Shift change improvement	1.5	-25%	node 10
p_{13}	Backlog's reduction	5	-30%	node 7
p_{14}	DMT reduction	2.5	-24%	node 5
p_{15}	Load site improvement	3	-25%	node 1

Preposition 1 If $X_u \cap X_v \neq \emptyset$ or $X_u \cup X_v = X_k \mid k \in \mathbf{S}$ Then $x_u + x_v \leq 1$.

In this case, X_k, X_u and X_v are the combinations sets associated with the variables x_k, x_u and x_v, respectively.

Let w_k be the cost of the combination project k, the first constraint limits the sum of cost to the budget available R. Moreover, the last constraint indicates that the decision variables must be binary.

$$\begin{aligned} \text{Maximize} \quad & \sum_{k=1}^{K} E[g_k] x_k \\ \text{Subject to} \quad & \sum_{k=1}^{q} w_k x_k \leq R \\ & x_u + x_v \leq 1 (\textit{ preposition } 1) \\ & x_k, x_u, x_v \in \{0,1\} \end{aligned} \tag{15}$$

For this study, it was available from the mine company a project portfolio with 15 candidate projects and a budget R of 15pu (per unit). Table 2 presents the projects to be

considered with the projects costs (c). Moreover, Table 2 shows the impacts of the projects and the process (or parameter) affected.

This problem was formulated according to Eq. 15 with a limit of 200 'relevant combinations'. The 'Gurobi Optimizer' software was used and the best project portfolio solution was found. Lastly, the optimal solution was converted in a set of isolate project (such as Table 2) and the expected gain of this solution was obtained using the queuing network model presented in this paper. As a result, the best portfolio was $\{1,0,0,0,0,1,1,1,1,0,1,0,0,0,0\}$, which provides a production increase of 1049.09 tonnes.

6 Conclusions

In the context of projects portfolio, there is no simple way to select the best portfolio when considering the interrelationship between projects. Thus, this paper showed a strategy based on the characteristics of the load-haulage cycle of an open-pit mine, that limits cohesively the interrelationships to be evaluated, based on an adaptation of the knapsack problem. Given that the computational time is a limited resource, an analytical approximation model was designed. The results showed that there is not significant difference between this model and DES simulation models which represents the same load-haulage cycle.

References

Jerry Banks. Introduction to Simulation. *Winter Sim. Conf.*, 2000.

Gunter Bolch, Stefan Greiner, Hermann de Meer, and Kishor S Trivedi. *Queueing Networks and Markov Chains: Modeling and Performance evaluation with Computer Science applications*. A Wiley-Interscience, 1998.

Jeffrey P Buzen. Computational Algorithms for Closed Queueing Networks with Exponential Servers. *Com. of ACM*, 16, 1973.

Christos G Cassandras. *Introduction to Discrete event Systems*. Springer Science Business Media, 2008.

Banu Y Ekren, Sunderesh S Heragu, Arvind Krishnamurthy, and Charles J Malmborg. An approximate solution for semi-open queueing network model of an autonomous vehicle storage and retrieval system. *Automation Science and Engineering, IEEE Trans. on*, 10(1):205–215, 2013.

S G Ercelebi and A Bascetin. Optimization of shovel-truck system for surface mining. *The Journal of The Southern African Institute of Mining and Metallurgy*, 109:433–439, 2009.

Michel C Fu. Optimization for Simulation: Theory vs. Practice. *INFORMS Journal on Computing*, 14:192–215, 2002.

W J Gordon and R R Newell. Closed Queueing Systems with Exponential Servers. *Oper. Research*, 15:254–265, 1967.

James R Jackson. Networks of waiting lines. *Operations research*, 5(4):518–521, 1957.

Raymond A Marie. Disappointments and Delights, Fears and Hopes induced by a few decades in Performance Evaluation. In *Perf. Eval. of Comp. and Communication Systems. Milestones and Future Challenges*, pages 1–9. Springer, 2011.

Barry L Nelson. Optimization via Simulation Over Discrete Decision Variables. *Tuts. in Oper. Research - INFORMS*, pages 193–207, 2010. doi:10.1287/educ.1100.0069.

C Dennis Pegden. Introduction to SIMAN. In *Proc. of the 15th conf. on Winter simulation-Volume 1*, pages 231–241. IEEE Press, 1983.

David Pisinger. A Minimal Algorithm for the Multiple-Choice Knapsack Problem. *European Journal of Oper. Research*, 83: 394–410, 1994. doi:10.1016/0377-2217(95)00015-I.

Cesar Monteiro Ribeiro. Modelagem e simulação de Sistemas a Eventos Discretos via redes de Petri estocásticas: Aplicação em mineração. Master's thesis, Universidade Federal de Minas Gerais, 2015.

Elmira Torkamani and Hooman Askari-Nasab. Verifying Short-Term Production Schedules using Truck-Shovel Simulation. *MOL Report Four*, pages 302:1–16, 2012.

Simulating the Effect of a Class of Sensor Fuzed Munitions for Artillery on a Multiple Target Element System

Henri Kumpulainen Bernt M. Åkesson

Information Technology Division, Finnish Defence Research Agency, Finland, bernt.akesson@mil.fi

Abstract

This paper presents a method for analyzing the effect of a class of artillery-launched sensor fuzed munitions on a target, which is a system consisting of several target elements and a fault logic. The target elements are armored vehicles and the munitions are designed specifically to attack single vehicles. We consider munitions which may contain one or two submunitions. We want to address the following questions: what is the probability of disabling the system given the number of ammunition, and similarly, how much ammunition is needed for disabling the system with a given confidence level. The proposed method is based on Markov chains rather than Monte Carlo simulation.

Keywords: Markov processes, probability, set theory, operations research, mathematical model, algorithms, weapons

1 Introduction

The term sensor fuzed munition is here used to denote munitions with target recognition capability and the ability to autonomously search for, detect, recognize and attack single target elements with specific signatures. Sensor fuzed munitions are generally intended for use against armored vehicles. Sensor fuzed munitions engage their targets from above from a distance using explosively formed projectiles (Dullum, 2008).

In this paper we consider sensor fuzed submunitions that are delivered by an artillery projectile. The projectile travels on a ballistic trajectory and ejects one or two sensor fuzed submunitions over a desired release point using a time fuze. The submunitions operate independently of each other, and after stabilizing and retarding their descent, scan the area beneath them for suitable target elements. Figure 1 illustrates the operation of the munitions.

A target consists of a number of target elements in a formation. Defeat of the target is defined by the damaging of specific combinations of target elements.

We propose an approach based on Markov chains. This approach has the benefit that once the effect of a single projectile has been calculated, the effects of further projectiles can be calculated by simple matrix operations.

Analytical models for different types of sensor fuzed munitions have been derived by (Halsør and Kvifte, 2003). This paper extends that work.

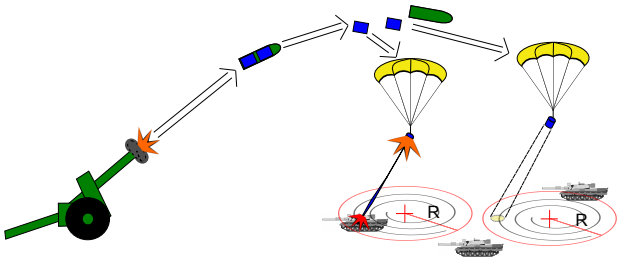


Figure 1. Schematic of the firing of an artillery projectile containing two sensor fuzed anti-tank submunitions.

A corresponding method as in this paper has been previously applied to fragmenting munitions by (Pettersson et al., 2011). The main difference is that a single fragmenting munition can damage all target elements in the area, whereas a sensor fuzed submunition can only damage a single target element.

An alternative approach that has been used before for this type of problems is to apply Monte Carlo methods, as outlined in e.g. (Halsør and Kvifte, 2003) and (NATO Standardization Office, 2012). Monte Carlo methods are known to be computationally expensive when we need accurate estimates of the kill probabilities, since their error is proportionate to $1/\sqrt{n}$ where n is the number of replications.

The paper is outlined as follows. First, an overview of the method is given. After this, a method for computing the probabilities of encountering different target element subsets is presented. Then the probabilities in the cases with one and two submunitions are derived. Finally, the failure probability of a system of target elements is considered.

2 Overview of the Method

An illustration of the method is provided in Figure 2. The computations are performed in three steps. First, the probabilities of a submunition encountering each subset of target elements are computed. After this, a state transition matrix is constructed and finally, the effect of a given number of munitions is computed. The state vector can be used together with fault logics to determine the probability of the target being broken.

The following assumptions are made:

1. The position of each target element is known

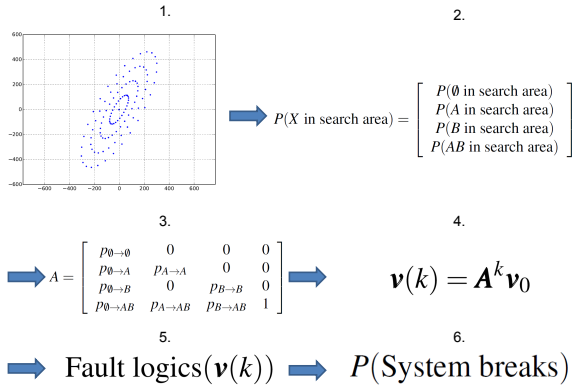


Figure 2. Illustration of the method.

2. The target elements are identical
3. A single submunition can only kill one target element
4. A target element has two states, functional and broken
5. Broken target elements are never attacked
6. The sensor is scanning, seeing only part of the search area at a time, and the munition will attack the first target it detects that fits given criteria. See Figure 1.
7. The search area, also known as the sensor footprint, is circular in shape
8. The target elements are stationary

The second assumption can be generalized by having different probabilities for detection, hit and kill for different target element types. The seventh assumption can be easily relaxed to other footprint shapes.

The carrier projectile is assumed to follow a ballistic trajectory, which can be computed using, e.g., a modified point-mass model. In this paper, the release point is for simplicity assumed to follow a bivariate normal distribution on the ground plane. The mean and standard deviations are assumed to be known.

A sensor fuzed submunition is characterized by the following parameters:

- Radius R of the search area
- Reliability p_{f}
- Detection probability p_{d}
- Hit probability p_{h}
- Kill probability $p_{\mathrm{k|h}}$, i.e., the probability of the target element breaking when hit

The probability of detection and the probability of kill may depend of the target element type together with weather and terrain conditions. The probability of detection may also depend on the distance from the center of the search area. In this paper the scanning method of the submunition is not specified, and thus all target elements inside the search area have equal probability of being detected first.

Some parameter values for existing sensor fuzed munitions can be found in open sources. For example, for a certain munition, (Dullum, 2008) reports a search area radius $R = 100$ m, and (Kosola and Solante, 2013) report a hit probability of $p_{\mathrm{h}} = 0.8$ and a kill probability of $p_{\mathrm{k|h}} = 0.95$.

3 Computing the Encounter Probabilities

We consider n identical target elements in the target area. Let us denote the set of target elements by $T = \{T_1, T_2, T_3, \ldots, T_n\}$, and the probability that the set of broken target elements after k artillery rounds is exactly X by $P_k(X)$.

Given a position (x,y) on the ground plane, there is a probability $p(x,y)$ of a sensor fuzed submunition starting its search in that point. The probability is directly the probability density function of the bivariate normal distribution. We look at a circular area with radius R around (x,y) and select all target elements within it. This corresponds to subset $X_i \subseteq T$. The integrand function returns a vector with $\dim(\boldsymbol{v}) = 2^n$, whose ith element has value $p(x,y)$.

Let $\boldsymbol{P}(\text{set in search area})$ be a vector of length 2^n containing the probabilities that a single round encounters a given subset of the target elements within its search area,

$$\boldsymbol{P}(\text{set in search area}) = \begin{bmatrix} P(\emptyset \text{ in search area}) \\ P(\{T_1\} \text{ in search area}) \\ P(\{T_2\} \text{ in search area}) \\ P(\{T_1, T_2\} \text{ in search area}) \\ \vdots \\ P(\{T_1, T_2, \ldots, T_n\} \text{ in search area}) \end{bmatrix}. \tag{1}$$

One approach, proposed in (Halsør and Kvifte, 2003), is to integrate over an area to obtain the probabilities of encountering each target element combination,

$$\boldsymbol{P}(\text{set in search area}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \boldsymbol{f}(x,y)\,\mathrm{d}x\mathrm{d}y \tag{2}$$

where the integrand $\boldsymbol{f}(x,y)$ is a vector-valued function, such that for $i = 1, \ldots, 2^n$

$$f_i(x,y) = \begin{cases} p(x,y), & \text{if subset } X_i \text{ is within distance } R \\ & \text{from point } (x,y) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Table 1. Encounter probabilities corresponding to the situation in Figure 3

Set X	*P(X in search area)*
$\emptyset$	0.0138
$\{A\}$	0.0018
$\{B\}$	0.0781
$\{A,B\}$	0.1427
$\{C\}$	0.1834
$\{A,C\}$	0.1521
$\{B,C\}$	0.0426
$\{A,B,C\}$	0.3853

where $p(x,y)$ is the value of the probability density function at point (x,y).

3.1 Example

The integral in (2) can be performed for an elliptical area centered on the aimpoint. Figure 3 illustrates the situation. The dispersion pattern is shown as ellipses. Three example functioning points are shown. At functioning point P_1 only target element B can be detected, at point P_2 A and C can be detected and at point P_3 all three target elements can be detected. When integrating over the area using (2) we obtain the encounter probabilities listed in Table 1.

4 State Transition Matrix

Let $P_k(X)$ be the probability that the set of target elements X is broken at time step k. Let $\boldsymbol{v}(k)$ be a vector that contains the values of $P_k(X)$ for every subset $X \subseteq T$ in bit order, i.e., $\boldsymbol{v}(k) = [P_k(\emptyset), P_k(\{T_1\}), P_k(\{T_2\}), P_k(\{T_1,T_2\}), \ldots]^T$, at time step k.

We can interpret the system as a state machine where a state is characterized by the set of broken target elements and during each time step exactly one round of the weapon system is fired. The vector $\boldsymbol{v}(k)$ contains the state probabilities. Initially, all target elements are functional, thus $\boldsymbol{v}(0) = [1,0,\ldots,0]^T$.

We can compute a state transition matrix $\boldsymbol{A}$ so that, for all k,

$$\boldsymbol{v}(k+1) = \boldsymbol{A}\boldsymbol{v}(k). \tag{4}$$

After firing k rounds the end state is given by

$$\boldsymbol{v}(k) = \boldsymbol{A}^k \boldsymbol{v}(0). \tag{5}$$

The state transition matrix $\boldsymbol{A}$ is defined as

$$\boldsymbol{A} = [P(X_j \to X_i)]_{1 \le i,j \le 2^n} \tag{6}$$

where $P(X_j \to X_i)$ is the probability of moving from state $X_j \subseteq T$ to state $X_i \subseteq T$.

Matrix $\boldsymbol{A}$ is rather large and has dimensions $2^n \times 2^n$. However, $\boldsymbol{A}$ is rather sparse, which means that all matrix operations will be faster when implemented using sparse matrices.

During an artillery firing an individual target element can either remain functional or break but broken target elements can never return back to functional. This means that $P(X_j \to X_i) = 0$ if $X_j \not\subseteq X_i$, and thus $P(X_j \to X_i)$ can only be nonzero if $X_j \subseteq X_i$. In (Pettersson et al., 2011) it was shown that when a single round is capable of killing all target elements the number of nonzero elements in $\boldsymbol{A}$ is at most 3^n.

If, on the other hand, at most n_K target elements can be killed by a single round, then $P(X_j \to X_i)$ can only be nonzero if $|X_i \setminus X_j| \le n_K$ and $X_j \cap X_i = X_j$. The number of nonzero elements matrix $\boldsymbol{A}$ is then

$$N(\text{nonzero elements in } \boldsymbol{A}) = \sum_{j=1}^{2^n} N(\text{nonzero elements in column } j \text{ of } \boldsymbol{A}) \le \sum_{i=0}^{n} \binom{n}{i} \sum_{m=0}^{n_K} \binom{n-i}{m}. \tag{7}$$

If one target element can be killed we have at most $(n+2)2^{n-1}$ nonzero elements and if two target elements can be killed the number of nonzero elements is at most $(n^2+3n+8)2^{n-3}$.

4.1 Projectile with a Single Sensor Fuzed Submunition

If the encountered set of target elements is X_k and the initial state is X_j, the set of killed target elements in the encountered set is $X_j \cap X_k$. The set of functional target elements in the encountered set is $X_k \setminus X_j$.

Since the target elements are assumed identical, for any target element $X \in X_k$ we have that

$$P(X \text{ killed} \mid X_k \text{ in search area and } X_j \cap X_k \text{ killed}) = P(1/|X_k \setminus X_j| \text{ killed}). \tag{8}$$

The probability of killing exactly one target element out of m is

$$P(1/m \text{ killed}) = (1-(1-p_\mathrm{d})^m)\, p_\mathrm{h} p_{\mathrm{k|h}} p_\mathrm{f}/m. \tag{9}$$

When only one target element can be killed each time step, the exact probability of moving from state $X_j \subseteq T$ to state $X_i \subseteq T$ in one time step is

$$P(X_j \to X_i) = P(X_j \to (X_j \cup \{T_i\})) = \sum_{\substack{X_k \subseteq T, \\ \text{s.t. } T_i \in X_k}} P(X_k \text{ in search area}) P(1/|X_k \setminus X_j| \text{ killed}). \tag{10}$$

Next we present an algorithm for calculating the state transition matrix more efficiently than the brute force approach of applying (10) for each nonzero element in the matrix.

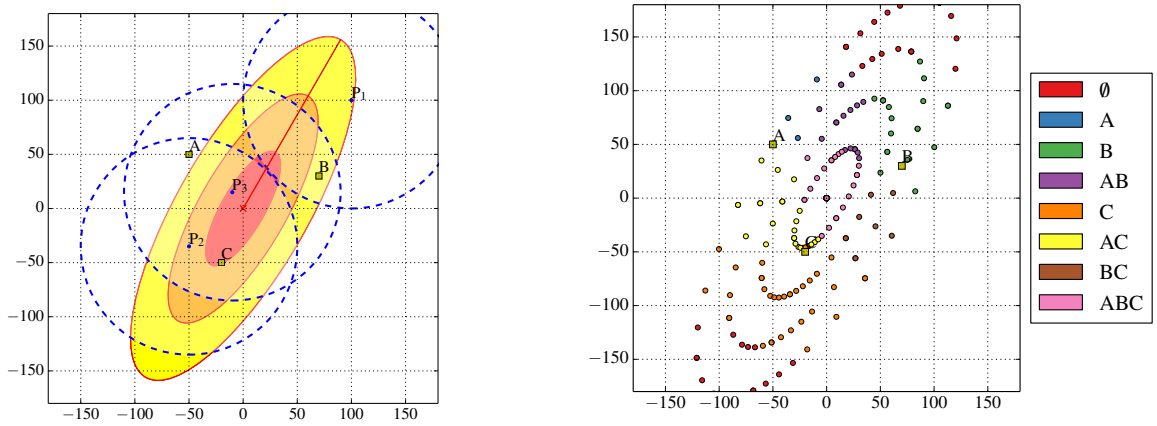


Figure 3. The figure on the left shows the positions of three target elements: A, B and C. The ballistic dispersion of the carrier projectile is marked by ellipses for one, two and three standard deviations. A single submunition is assumed. Three example functioning points P_1, P_2 and P_3 have been plotted, with the sensor footprints outlined with dashed circles. The figure on the right shows the subsets found in each integration point.

Let us define $Q(X)$ as the probability that X is a subset of the set of target elements inside the search area. We can calculate this probability as

$$Q(X) = \sum_{\substack{Y \subseteq T, \\ \text{s.t. } X \subseteq Y}} P(Y \text{ in search area}). \tag{11}$$

The probability that a set of target elements X is inside the search area, but target element T_i is not, is $Q(X) - Q(X \cup \{T_i\})$. Applying this recursively we can rewrite $P(X \text{ in search area})$ in terms of Q as

$$P(X \text{ in search area}) = \sum_{i=0}^{|T \setminus X|} \sum_{\substack{Y \subseteq (T \setminus X), \\ \text{s.t. } |Y|=i}} \begin{cases} Q(X \cup Y), & i \text{ is even} \\ -Q(X \cup Y), & i \text{ is odd} \end{cases} \tag{12}$$

Unlike $P(X_k \text{ in search area})$, which also depends on the total number of target elements in the search area, the value of $Q(X_k)$ does not change when the number of target elements in $T \setminus X_k$ changes. If we write (10) in terms of Q, we can disregard the broken target elements in X_j, reducing the number of subsets to take into account. The probability that a target element T_i is killed is then

$$P(X_j \to (X_j \cup \{T_i\})) = \sum_{\substack{X_k \subseteq T, \\ \text{s.t. } T_i \in X_k}} P(X_k \text{ in search area}) P(1/|X_k \setminus X_j| \text{ killed}) =$$

$$\sum_{\substack{X_k \subseteq (T \setminus X_j), \\ \text{s.t. } T_i \in X_k}} \sum_{i=0}^{|T \setminus (X_k \cup X_j)|} \sum_{\substack{Y \subseteq (T \setminus (X_k \cup X_j)), \\ \text{s.t. } |Y|=i}} \begin{cases} Q(X_k \cup Y) P(1/|X_k| \text{ killed}), & i \text{ is even} \\ -Q(X_k \cup Y) P(1/|X_k| \text{ killed}), & i \text{ is odd} \end{cases} \tag{13}$$

Since all possible $X_k \cup Y$ are equivalent to some value of X_k and the value of $P(1/|X_k| \text{ killed})$ does not depend on Y, the sum can be factorized as

$$P(X_j \to X_j \cup \{T_i\})) = \sum_{\substack{X_k \subseteq (T \setminus X_j), \\ \text{s.t. } T_i \in X_k}} Q(X_k) \sum_{i=0}^{|X_k|-1} \begin{cases} \binom{|X_k|-1}{i} P(1/(|X_k| - i) \text{ killed}), & i \text{ is even} \\ -\binom{|X_k|-1}{i} P(1/(|X_k| - i) \text{ killed}), & i \text{ is odd} \end{cases} \tag{14}$$

$$K(n) = \sum_{i=0}^{n-1} \begin{cases} \binom{n-1}{i} P(1/(n-i) \text{ killed}), & i \text{ is even} \\ -\binom{n-1}{i} P(1/(n-i) \text{ killed}), & i \text{ is odd} \end{cases} \tag{15}$$

We can define a recursive function

$$f(X,Y) = \begin{cases} f(X \setminus \{T_i\}, Y) + f(X \setminus \{T_i\}, Y \cup \{T_i\}), & T_i \in X \\ Q(Y) K(|Y|), & X = \emptyset \end{cases} \tag{16}$$

which has the property

$$f(T \setminus (X_j \cup \{T_i\}), \{T_i\}) = \sum_{\substack{X_k \subseteq (T \setminus X_j), \\ \text{s.t. } T_i \in X_k}} Q(X_k) K(|X_k|) = P(X_j \to X_j \cup \{T_i\})). \tag{17}$$

We notice that by selecting X and Y appropriately, $f(X,Y)$ gives us the nonzero nondiagonal elements of the state transition matrix. All the relevant values of K, Q and $f(X,Y)$ can be precalculated in order to create each element of the matrix in constant time, as demonstrated in the algorithm in Fig. 4. The overall complexity of this algorithm is $O(n^2 \cdot 2^n)$.

Input: $T, n = |T|, \boldsymbol{P}(\text{set in search area}), \boldsymbol{P}(1/m \text{ killed})$
  for $i = 1$ **to** n **do** {Precalculate $K(i)$ to table K_i}
    $b_j \leftarrow \begin{cases} 1, & j = 1 \\ a_j - a_{j-1}, & j = 2, \ldots, i-1 \\ -1, & j = i \text{ and } i \text{ is even} \\ 1, & j = i \text{ and } i \text{ is odd} \end{cases}$
    $K_i \leftarrow \sum_{j=1}^{i} b_j P(1/(i-j+1) \text{ killed})$
    $\boldsymbol{a} \leftarrow \boldsymbol{b}$
  end for
  for $k = 0$ **to** n **do** {Precalculate $Q(X)$ to hash table Q_X}
    for all subsets X such that $X \subseteq T$ and $|X| = n - k$ **do**
      $Q_X \leftarrow P(X \text{ in search area})$
      $S_{X,1} \leftarrow P(X \text{ in search area})$
      for all elements T_i such that $T_i \in T \setminus X$ **do**
        $Q_X \leftarrow Q_X + S_{X \cup T_i, i}$
        $S_{X,i+1} \leftarrow S_{X,i} + S_{X \cup T_i, i}$
      end for
    end for
  end for
  for $k = 1$ **to** n **do** {Precalculate $f(X,Y)$ to hash table $f_{X,Y}$}
    for all subsets Y such that $Y \subseteq T$ and $|Y| = k$ **do**
      $X \leftarrow \emptyset$
      $f_{\emptyset,Y} \leftarrow Q_Y \cdot K_{|Y|}$
      Move the last element of Y from Y to X
      for $j = 2$ **to** k **do**
        Move the last element of Y from Y to X
        for $i = 1$ **to** j **do**
          $X_m \leftarrow \begin{cases} X_1, & i > 1 \\ X_2, & i = 1 \end{cases}$
          Move element X_i from X to Y
          $f_{X,Y} \leftarrow f_{X \setminus \{X_m\}, Y} + f_{X \setminus \{X_m\}, Y \cup \{X_m\}}$
          Move element X_i back from Y to X
        end for
      end for
    end for
  end for
  {Construct the state transition matrix $\boldsymbol{A} = [a_{mj}]_{1 \le m,j \le 2^n}$}
  $\boldsymbol{A} \leftarrow 0_{2^n, 2^n}$
  for all subsets X_j such that $X_j \subseteq T$ **do**
    $a_{jj} \leftarrow 1$
    for all elements T_i such that $T_i \in (T \setminus X_j)$ **do**
      Get row number m such that $X_m = X_j \cup \{T_i\}$
      $a_{mj} \leftarrow f_{T \setminus (X_j \cup \{T_i\}), \{T_i\}}$
      $a_{jj} \leftarrow a_{jj} - a_{mj}$
    end for
  end for
  return $\boldsymbol{A}$

Figure 4. Algorithm for creating state transition matrix for a single sensor fuzed submunition.

4.2 Example

With two target elements, A and B, we have state vector $\boldsymbol{v}$

$$\boldsymbol{v} = \begin{bmatrix} P(\emptyset \text{ killed}) \\ P(\text{A killed}) \\ P(\text{B killed}) \\ P(\text{AB killed}) \end{bmatrix} \tag{18}$$

and state transition matrix $\boldsymbol{A}$

$$\boldsymbol{A} = \begin{bmatrix} p_{\emptyset \to \emptyset} & 0 & 0 & 0 \\ p_{\emptyset \to A} & p_{A \to A} & 0 & 0 \\ p_{\emptyset \to B} & 0 & p_{B \to B} & 0 \\ 0 & p_{A \to AB} & p_{B \to AB} & 1 \end{bmatrix} \tag{19}$$

where

$$\begin{aligned} p_{\emptyset \to A} &= f(B,A) = f(\emptyset,A) + f(\emptyset,AB) \\ &= Q(A)K(1) + Q(AB)K(2) && (20) \\ p_{\emptyset \to B} &= f(A,B) = f(\emptyset,B) + f(\emptyset,AB) \\ &= Q(B)K(1) + Q(AB)K(2) && (21) \\ p_{A \to AB} &= f(\emptyset,B) = Q(B)K(1) && (22) \\ p_{B \to AB} &= f(\emptyset,A) = Q(A)K(1) && (23) \\ p_{\emptyset \to \emptyset} &= 1 - p_{\emptyset \to A} - p_{\emptyset \to B} && (24) \\ p_{A \to A} &= 1 - p_{A \to AB} && (25) \\ p_{B \to B} &= 1 - p_{B \to AB} && (26) \end{aligned}$$

and

$$\begin{aligned} Q(A) &= P(A \text{ in search area}) + P(AB \text{ in search area}) && (27) \\ Q(B) &= P(B \text{ in search area}) + P(AB \text{ in search area}) && (28) \\ Q(AB) &= P(AB \text{ in search area}) && (29) \\ K(1) &= \binom{0}{0} P(1/1 \text{ killed}) && (30) \\ K(2) &= \binom{1}{0} P(1/2 \text{ killed}) - \binom{1}{1} P(1/1 \text{ killed}) && (31) \end{aligned}$$

4.3 Projectile with Two Sensor Fuzed Submunitions

We now consider the case where an artillery projectile contains two sensor fuzed submunitions, which are released at some altitude. The release point is still assumed to follow a bivariate normal distribution. The center of the search areas of each submunition is assumed to be located at given distances from the release point in the direction of fire, since the submunitions will have forward motion before being sufficiently slowed down, as illustrated in Figure 5.

For tube artillery projectiles containing two submunitions, the centers of the search areas of the submunitions are separated by approximately 300 m, according to (Dullum, 2008). If the radius of the search area for each

submunition is 100 m, the search areas will not overlap. Therefore, in this paper the search areas are assumed to be non-overlapping. The submunitions are also assumed identical.

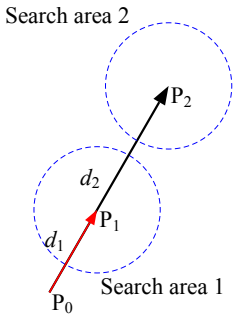


Figure 5. Sensor fuzed munition containing two indentical submunitions, whose non-overlapping search areas are outlined with dashed circles. The release point is at P_0, after which submunition 1 travels distance d_1 and submunition 2 distance d_2. The direction of fire is represented by the vectors.

The encounter probabilities have to be constructed for both search areas. We denote the subset in the first search area by S_1 and the subset in the second search area by S_2, with the constraint that $S_1 \cap S_2 = \emptyset$, i.e., the search areas contain disjoint subsets. The probability vector $\boldsymbol{P}$(sets in search area) can be calculated in a similar manner as in the single submunition case and will have length 3^n. The state transition matrix can be constructed using an approach similar to the single submunition case. Let us redefine the variables as

$$Q(X,Y) = \sum_{\substack{S_1 \subseteq T, \\ \text{s.t. } X \subseteq S_1}} \sum_{\substack{S_2 \subseteq T \setminus S_1, \\ \text{s.t. } Y \subseteq S_2}} P(S_1, S_2 \text{ in search area}) \quad (32)$$

$$P(0/n \text{ killed}) = \begin{cases} 1 - n \cdot P(1/n \text{ killed}), & n > 0 \\ 1, & n = 0 \end{cases} \quad (33)$$

$$K_0(n) = \begin{cases} \sum_{i=0}^{n} \binom{n}{i} P(0/(n-i) \text{ killed}), & i \text{ is even} \\ -\sum_{i=0}^{n} \binom{n}{i} P(0/(n-i) \text{ killed}), & i \text{ is odd} \end{cases} \quad (34)$$

$$f_1(X,Y,Z) = \begin{cases} f_1(X \setminus \{T_i\}, Y, Z) + f_1(X \setminus \{T_i\}, Y \cup \{T_i\}, Z) \\ \quad + f_1(X \setminus \{T_i\}, Y, Z \cup \{T_i\}), & T_i \in X \\ Q(Y,Z)K(|Y|)K_0(|Z|) \\ \quad + Q(Z,Y)K(|Z|)K_0(|Y|), & X = \emptyset \end{cases} \quad (35)$$

$$f_2(X,Y,Z) = \begin{cases} f_2(X \setminus \{T_i\}, Y, Z) + f_2(X \setminus \{T_i\}, Y \cup \{T_i\}, Z) \\ \quad + f_2(X \setminus \{T_i\}, Y, Z \cup \{T_i\}), & T_i \in X \\ (Q(Y,Z) + Q(Z,Y))K(|Y|)K(|Z|), & X = \emptyset \end{cases} \quad (36)$$

$$\begin{aligned} P(X_j \to X_j \cup \{T_1\}) &= f_1(T \setminus (X_j \cup \{T_1\}), \{T_1\}, \emptyset) \\ P(X_j \to X_j \cup \{T_1, T_2\}) &= f_2(T \setminus (X_j \cup \{T_1, T_2\}), \\ &\qquad \{T_1\}, \{T_2\}) \end{aligned} \quad (37)$$

With the above definitions, Q, f_1 and f_2 can be precalculated in the same way as Q and f in Fig. 4. The state transition matrix for two submunitions can be created in time $O(n \cdot 3^n)$, which is the complexity of precalculating $Q(X,Y)$ recursively for all possible $X, Y \in T$.

5 Failure Probability of the System

Arbitrary fault logics may be used to determine which broken subsets correspond to a broken system. The probability that the system breaks is the sum of these subset destruction probabilities in $\boldsymbol{v}(k)$,

$$P(\text{system breaks}) = \sum_{\substack{X_i \subseteq T, \\ \text{s.t. } g(X_i)=1}} v_i(k) \quad (38)$$

where function g contains the fault logics. The value of $g(X)$ is 1 if X corresponds to a broken system according to the fault logic and 0 otherwise.

The smallest number of projectiles needed to guarantee failure of the system with a given confidence level α can be obtained simply by starting at $k = 1$ and calculating P(system breaks) repeatedly while increasing the value of k until $P(\text{system breaks}) \geq \alpha$.

6 Conclusions

The placement of the aimpoint has a significant impact as well as the trajectories of the submunitions, here modeled as offset vectors. The trajectories of the submunitions from expulsion from the carrier projectile to their activation need further study.

We also need to address the target location error and the mean release point error. The mean release point may not necessarily coincide with the actual aimpoint and the aimpoint may also have a systematic error. One approach to handling these errors is to assume that the errors follow a bivariate normal distribution and discretize the region into a number of points and iterating over them.

The method can be generalized to having target elements with different properties, which would translate to different detection, hit and kill probabilities.

Acknowledgment

The authors would like to thank Ilmari Kangasniemi and Janne Valtonen for their valuable input on the manuscript.

References

Ove Dullum. Cluster weapons – military utility and alternatives. FFI-rapport 2007/02345, Norwegian Defence Research Establishment (FFI), Norway, 2008.

Marius Halsør and Lars Kvifte. Metoder for effektberegning av smart artilleriammunisjon [Methods for determining

the effect of smart artillery ammunition]. FFI/RAPPORT-2003/00084, Forsvarets Forskningsinstitutt (Norwegian Defence Research Establishment), 2003.

Jyri Kosola and Tero Solante. *Digitaalinen taistelukenttä - informaatioajan sotakoneen tekniikka [Digital Battlefield]*, volume 35 of *Julkaisusarja 1*. Maanpuolustuskorkeakoulu, Sotatekniikan laitos, Helsinki, 3rd edition, 2013. ISBN 978-951-25-2503-4.

NATO Standardization Office. STANAG 4654: Indirect fire appreciation modelling. Standardization agreement, NATO Standardization Office, Brussels, 2012.

Ville Pettersson, Eric Malmi, Sampo Syrjänen, Bernt Åkesson, Tapio Heininen, and Esa Lappi. Simulating the effect of indirect fire on a multiple target element system. In Sergey Repin, Timo Tiihonen, and Tero Tuovinen, editors, *Proceedings of CAO2011: ECCOMAS thematic conference on computational analysis and optimization*, volume No. A 1/2011 of *Reports of the Department of Mathematical Information Technology, Series A. Collections*, Jyväskylä, Finland, June 9–11 2011. University of Jyväskylä. ISBN 978-951-39-4331-8.

Simulation Environment for Development of Unmanned Helicopter Automatic Take-off and Landing on Ship Deck

Antonio Vitale[1] Davide Bianco[1] Gianluca Corraro[1] Angelo Martone[1] Federico Corraro[1]
Alfredo Giuliano[2] Adriano Arcadipane[2]

[1]On-boar Systems and ATM Department, CIRA - Italian Aerospace Research Centre, Capua (CE), Italy, {a.vitale, d.bianco, g.corraro, a.martone, f.corraro}@cira.it

[2]Electrical and Avionics Systems Department, Finmeccanica - Helicopter Division, Cascina Costa (VA), Italy, {Alfredo.Giuliano, Adriano.Arcadipane}@finmeccanica.com

Abstract

Helicopter take-off and landing operations on ship carrier are very hazardous and training intensive. Guidance, Navigation and Control algorithms can help pilots to face these tasks by significantly reducing the workload and improving safety level. Anyway, the design and verification of such algorithms require the availability of suitable simulation environments that shall be a trade-off between simplicity and accuracy. This paper presents the simulation models developed to support the design, pre-flight verification and validation of helicopter trajectory generation and tracking algorithms for automated take-off and landing on a frigate deck. The process for generation and testing of the code to be integrated into the real-time Software-In-the-Loop simulator is also described. Such fast time and real-time simulation environments contributed to reduce algorithms design time, risks and costs, by limiting the required flight test activities. Take-off and landing algorithms developed by using the proposed simulation environments were successfully demonstrated in flight.

Keywords: GNC, helicopter, sensor, ship, turbulence

1 Introduction

Vertical take-off and landing operations of aerial vehicle on ship's deck enhance mission capabilities for military and civilian users. Anyway, these operations are the most dangerous flight phases for helicopters (Padfield, 1998; Lee, 2005). Indeed, a pilot have to deal with an invisible ship air wake, poor visible cueing and a landing spot which is heaving, rolling, pitching and yawing. At the same time the pilot shall also monitor vehicle's structural, aerodynamic and control limits. Moreover, operations take place in close proximity to the superstructure of the ship, that means there is little margin for error and the consequences of a significant loss of positional accuracy by the pilot can be severe.

The availability of Guidance Navigation and Control (GNC) algorithms for automatic operations can help pilots to face these tasks by significantly reducing operator workload, improving safety level and flight handling qualities. To develop these algorithms, suitable simulation environments are essential in order to reduce the flight test time and cost and to establish safe operating envelopes. The simulation tools shall be able to model all the relevant phenomena, such as helicopter flight dynamics (including on board sensors and actuators), the motion of the ship for the given sea state, the influence on the helicopter of the ship air wake and of the environment in general.

It is worth to note that modelling and simulation of each of the above listed phenomena is not a trivial task. Indeed, the simulation of the helicopter flight behavior includes kinematics, dynamics and aerodynamics of its subsystems (main rotor, fuselage, empennage, tail rotor, power plant, primary flight control system, on board sensors).

The vehicle's equations of motion, even if presented in several textbooks (Padfield, 1996; Johnson, 1994), are differential high order, nonlinear, coupled, and contain a large number of parameters, which often cannot be directly measured (Tishler et al., 2006). On the other hand, simplified models, which are able to catch the relevant dynamics, are typically required for GNC design purpose (Lee, 2005), to enhance physical understanding and lower the computational load. To this aim, linear parametrized models have been widely used (Tishler et al., 2006), but they are inadequate for accurate simulation of the vehicle dynamics when state variables significantly deviate from the linearization point (Gavrilets, 2006). Therefore, a suitable trade-off between model complexity and simulation accuracy shall be performed.

Another relevant topic concerns ship motion, which is an important issue for helicopter deck operations. For helicopter GNC algorithms design and analysis purpose, ship motion is usually represented through linear models or simplified nonlinear models with benign nonlinearities to capture the essential behavior of the vessel (Li, 2009). Ship motion can also be modelled using pre-computed or measured time histories (Carico et al., 2003). In any case, the ship model shall take into account the effect of the environment, and, in particular,

of the sea waves (Perez, 2005), which lead an undesirable low frequency disturbance into the motion of the vessel.

Finally yet importantly, the ship produces an air wake, which affects the helicopter dynamics. Indeed, ship air wake contains large velocity gradients and area of turbulence, generated by complex mechanisms of vortex dynamics near the ship deck, which greatly impair controllability of the flying vehicle and require additional control efforts to avoid accidents and to compensate abrupt changes in thrust level. Several accurate and complex CFD models of ship air wake are proposed in the literature (Kääriä, 2012), but CFD simulations produce a large amount of data and their use for GNC design and real time testing is usually unfeasible (Lee, 2005).

Stochastic turbulence models have been also proposed to represent the air wake with reasonable accuracy (Lee, 2005; Yang et al., 2009). These models may provide some insight into the effects of the air wake that are typically enough relevant for real-time simulations and flight control systems design.

It is also worth to note that, with reference to all the discussed models, a suitable code generation procedure and testing methodology shall be defined, in order to generate reliable real-time simulation models, applicable for GNC algorithms verification and performance assessment.

This paper presents the simulation environment developed by the Italian Aerospace Research Centre and Finmeccanica in order to support the design and verification of algorithms for helicopter trajectory generation and tracking, during an automated take-off and landing on a frigate deck. Matlab/Simulink was used for implementing such simulation environment, which constitutes an alternative to the already existing Finmeccanica GNC validation environment.

The proposed models, although simplified, are able to take into account the main effects of the sea's disturbance on the ship motion and of the ship air wake on the helicopter trajectory. Concerning the helicopter vehicle dynamics, its model emulates the relevant closed loop performance of the vehicle and includes operating envelope limitations through a model for aerodynamic forces and thrust computation, whose parameters are identified from experimental data.

The paper also includes some fast time simulation results compared to experimental data, demonstrating that the proposed simulation environment is accurate enough for GNC algorithms design.

Finally, some models of the above mentioned simulation environment were also integrated into the detailed Software-in-the-Loop Simulator of Finmeccanica, to perform real-time verification and validation of the whole Flight Management System (FMS). Therefore, a real-time automatic code generation process has been defined and implemented, in order to keep consistency between the simulation environment used for design, and the one used for final software verification. The paper briefly describes such generation process, which allowed producing reliable software code, compliant to DO-178C standard and Finmeccanica own implementation rules.

The proposed fast time simulation environment dramatically reduced the algorithms design time, risks and costs, by limiting the required flight test activities.

The take-off and landing algorithms developed by using the simulation environment described in this paper were successfully demonstrated in flight, by means of a full-size optionally piloted helicopter: the Finmeccanica SW-4 SOLO.

2 Simulation Models

The model based design process of a Guidance Navigation and Control system requires the development of simulation models with different complexity level, to be used in the various development phases, as shown in Figure 1.

The present section describes the mathematical models integrated into the simulation environment that was employed to design the helicopter trajectory generation and tracking algorithms for automated take-off and landing on a frigate deck. Figure 2 shows the functional architecture of such environment.

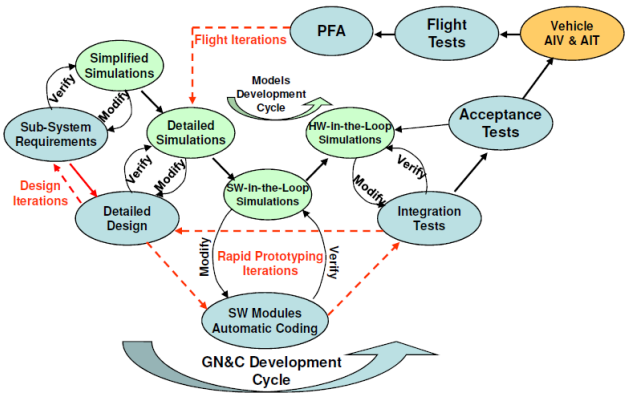


Figure 1. GNC Technology Development Cycle.

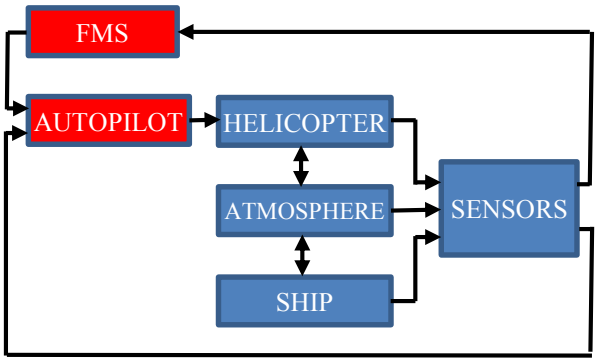


Figure 2. Simulation Environment functional architecture.

The blue blocks represent the simulation models, scope of this paper, whereas the red blocks are the GNC algorithms. The following sub-sections describe in detail each blue block.

2.1 Helicopter Model

Trajectory generation and tracking algorithms typically require the knowledge of vehicle's position and velocity only. Therefore, for the design and preliminary verification of such algorithms, it is sufficient and cost effective to model only the closed loop attitude dynamic response of the vehicle coupled with the high-level modes of the autopilot system. With this approach, a rigid body with three degrees of freedom, subject to external forces, and the rotational dynamic response of the vehicle to the autopilot commands represent the helicopter dynamics.

While this modelling approach is widely used in fixed-wing aircraft for guidance algorithms design, it is quite unusual for helicopters, because it does not take into account the coupled dynamics of the rotor flexibility with the helicopter rigid flight mechanics (Tishler et al., 2006).

Key original contribution of this paper is the development of a mixed empirical and physical formulation of the equations, so that the resulting simulation model includes only the low frequency effects of the neglected helicopter dynamics. As demonstrated by the comparisons with flight data reported in this paper, this allows obtaining enough accurate simulation results during the quasi-static manoeuvers of take-off and landing, while still taking into account the disturbance effects of wind and ship air wake.

The model assumes flat and fixed Earth, with constant gravity acceleration, and quasi-stationary variation of the vehicle mass (only due to fuel consumption).

The model's commands are the reference attitude and collective, while wind velocity ($\underline{V}_W$) is the disturbance input.

The actual attitude (φ_H, ϑ_H, ψ_H) and collective (δ_{coll}) of the vehicle, used in (1) for computation of forces, are modelled by unitary gain second order filters applied to the commands provided as input to the helicopter model. Such filtered Euler angles and collective and their rates are also saturated to account for actuator velocity limitations and some inner autopilot protection functions. Overall, the linear filters and related saturations model the closed loop performance of the inner autopilot modes. The parameters of both these filters and saturations are scheduled with respect to airspeed and they were identified by analyzing flight data gathered in specific manoeuvers.

The outputs are the helicopter position, velocity, load factors, actual attitude and angular rates.

The following equations of motion of the vehicle centre of mass (CoM), in North-East-Down (NED) inertial reference frame (McCormick, 1995), compute such outputs:

$$m\dot{\underline{V}} = \underline{F}\left(\varphi_H, \vartheta_H, \psi_H, \delta_{coll}, \underline{V}_W\right) \quad (1)$$

$$\dot{\underline{P}}_H = \underline{V}_H \quad (2)$$

$$\dot{h} = -w \quad (3)$$

where $\underline{V}$ is the inertial velocity vector, $\underline{V}_H$ and w are its horizontal (included into the North-East plane) and vertical components (positive down), respectively; $\underline{P}_H$ is the horizontal position and h the altitude of the vehicle CoM; m is the helicopter mass and $\underline{F}$ is the resultant force vector acting on the vehicle.

The force vector $\underline{F}$ is composed by gravitational force $\underline{W}$ (constant, and directed along the down axis of the NED reference frame), aerodynamic force $\underline{F}_A$ and propulsive force $\underline{T}$.

The computation of aerodynamic and thrust forces is first performed in the vehicle body reference frame, and then it is rotated in NED reference frame. The aerodynamic forces in body axes (F_{Ai}^B) are as follows:

$$F_{Ai}^B = q_{dyn}\sum_j S_j c_{i,j}(\alpha,\beta) \quad (4)$$

$$V_{TAS}^2 = \left(\underline{V}-\underline{V}_W\right)^T\cdot\left(\underline{V}-\underline{V}_W\right) \quad (5)$$

$$q_{dyn} = 0.5\rho\cdot V_{TAS}^2 \quad (6)$$

$$\alpha = tg^{-1}\left(w_{TAS}/u_{TAS}\right) \quad (7)$$

$$\beta = \sin^{-1}\left(v_{TAS}/V_{TAS}\right) \quad (8)$$

where q_{dyn} is the dynamic pressure, ρ is the air density, $V_{TAS} \equiv (u_{TAS}, v_{TAS}, w_{TAS})$ is the helicopter true airspeed, S_j is the reference aerodynamic surface of the j-th aerodynamic component (that is, fuselage, vertical and horizontal stabilizers) and $c_{i,j}$ the corresponding aerodynamic non-dimensional coefficient, which depends on the angle of attack α and sideslip angle β. Tabled functions express the aerodynamic coefficients using data extrapolated from flight experiments.

It is worthy to note that the aerodynamic angles α and β are not defined when the helicopter airspeed is null, for example in hover condition with null wind speed. In this case, the aerodynamic forces are negligible and the aerodynamic angles are not computed.

The propulsive force is evaluated by using the following semi-empirical linear model (Gavrilets, 2003):

$$T = z_{coll}\left(V_{TAS}\right)\cdot\delta_{coll} + z_w\left(V_{TAS}\right)\cdot w \quad (9)$$

The parameter z_{coll} is a gain between the thrust and the collective command δ_{coll} in level flight trim conditions. It is scheduled as a function of the forward speed of the aircraft with respect to air, and its values were identified applying a best-fit procedure of the rotor thrust data in

different flight conditions provided by the helicopter manufacturer.

The parameter z_w relates the thrust to the vertical speed. Although an analytical relation exists to express these parameters as function of vehicle characteristics (Gavrilets, 2003), in the present work, z_w was computed by fitting experimental data in climb and descent flight, and it is expressed as fraction of z_{coll}.

The thrust vector is assumed to point in the opposite direction of the body Z-axis. This hypothesis allows reproducing in simulation the trim values of pitch angle experimented in flight by the vehicle in level flight conditions.

2.2 Atmosphere Model

This model is in charge to reproduce the environmental conditions, in which the helicopter flies, that can influence the vehicle behaviour.

The model includes computation of atmospheric parameters (air density and temperature, static and dynamic pressure), wind velocity (wind shear, wind gust, atmospheric turbulence), and ship air wake experimented by the helicopter, based on its current position and velocity. International Standard Atmosphere (McCormick, 1995), von Karman model (von Karman, 1948) and standard wind model (MIL-F-8785C, 1991) are used for atmospheric parameters, turbulence and wind shear and gust, respectively.

Another element of originality included in this paper concerns the simplified ship air wake model, which is implemented as a stochastic phenomenon through a parameter modification of the von Karman turbulence model (von Karman, 1948).

In this model, independent white noise processes are suitably filtered to yield the desired forms of output power spectral density. The transfer functions (X_{ug}, X_{vg}, X_{wg}) of these linear filters in the Laplace domain are:

$$X_{ug}(s) = \sigma_u \sqrt{(2 \cdot L_u)/(\pi \cdot V_{TAS})} \cdot \frac{1}{1 + (L_u / V_{TAS}) \cdot s} \tag{10}$$

$$X_{vg}(s) = \sigma_v \sqrt{(2 \cdot L_v)/(\pi \cdot V_{TAS})} \cdot \frac{1 + 2\sqrt{3} \cdot (L_v / V_{TAS}) \cdot s}{\left[1 + (2 \cdot L_v)/V_{TAS} \cdot s\right]^2} \tag{11}$$

$$X_{wg}(s) = \sigma_w \sqrt{(2 \cdot L_w)/(\pi \cdot V_{TAS})} \cdot \frac{1 + 2\sqrt{3} \cdot (L_w / V_{TAS}) \cdot s}{\left[1 + (2 \cdot L_w)/V_{TAS} \cdot s\right]^2} \tag{12}$$

where σ_u, σ_v, σ_w and L_u, L_v, L_w are gains and scale factors, respectively, to be tuned through the analysis of CFD or experimental (wind tunnel or flight test) data.

Anyway, due to the unavailability of these data, such model parameters and their dependencies from helicopter state variables were determined through literature analysis and physical considerations.

The scale factors are set proportional to the characteristic lengths of the ship super-structure, which generates the wake. Since the effects of the wake on helicopter depend also from ship-helicopter relative position, the filters gains varied linearly with the ratio between ship speed and the square of the helicopter-ship distance.

Moreover, to take into account the local effect of the air ship wake disturbance and its dependence on wind direction, the wake's perturbation is only active within a limited size parallelepiped, which is oriented parallel to the wind speed and has width equal to the section of the super-structure orthogonal to the wind direction, length equal to three times the superstructure's section parallel to the wind direction, and height equal to three times the superstructure's height.

2.3 Ship Model

The ship translational motion is represented through kinematic relations, for the computation of undisturbed centre of mass position and velocity, plus an additive stochastic model, which simulates the sea wave disturbance on the ship. The applied equations for nominal position and velocity computation are:

$$\dot{\underline{V}}_{No} = \begin{bmatrix} a_x & a_y & 0 \end{bmatrix}^T \tag{13}$$

$$\dot{\underline{P}}_{No} = \begin{bmatrix} u_{No} & v_{No} & -w_{No} \end{bmatrix}^T \tag{14}$$

where a_x and a_y are the commanded horizontal acceleration of the ship; $\underline{V}_{No} \equiv (u_{No}, v_{No}, w_{No})$ and $\underline{P}_{No} \equiv (x_{No}, y_{No}, h_{No})$ are nominal velocity and position in NED reference frame, respectively. The actual position $\underline{P}_N \equiv (u_N, v_N, w_N)$ and velocity $\underline{V}_N \equiv (x_N, y_N, h_N)$ are calculated by adding the sea disturbance $\underline{\eta} \equiv (\eta_x, \eta_y, \eta_z)$ to nominal values:

$$\underline{V}_N = \begin{bmatrix} u_{No} & v_{No} & 0 \end{bmatrix}^T + \dot{\underline{\eta}} \tag{15}$$

$$\underline{P}_N = \underline{P}_{No} + \underline{\eta} \tag{16}$$

The attitude equations are defined independently as follows:

$$\begin{bmatrix} \varphi_N \\ \theta_N \\ \psi_N \end{bmatrix} = \begin{bmatrix} \varphi_0 \\ \theta_0 \\ \psi_0 \end{bmatrix} + \begin{bmatrix} \eta_\varphi \\ \eta_\theta \\ \eta_\psi \end{bmatrix} \tag{17}$$

It is assumed that the ship does not steer when helicopter is close, therefore its Euler angles (φ_N, θ_N, ψ_N) only depend on their initial values and on the sea disturbance (η_φ, η_θ, η_ψ).

The stochastic variables introduced in (16) and (17) for representing the sea disturbance are generated using the same equations. A mean velocity V_S and displacement D_S produced by the disturbance is associated to each variable. V_S and D_S depend on ship speed and sea state, and are provided by look up tables, which collect experimental data.

The sea disturbance is periodic and it pulsation ω is given by (Holthuijsen, 2017)

$$\omega = 2\pi V_S / D_S \quad (18)$$

The time evolution of the generic component of the sea disturbance η_i is then evaluated as follows

$$\eta_i(t) = A_i \sin(\omega_i t) \quad (19)$$

The gain A_i is a random variable with Rayleigh distribution, whose parameters depend on sea state and ship speed and are provided by a look up table based on experimental observations. During a simulation, the gain A_i is updated at the end of each wave period (that is, each D_S / V_S seconds) by performing a new random draw.

2.4 Sensor Model

Two kinds of sensors are available on-board the helicopter and are included into the simulation environment: a standard navigation suite (composed of an inertial navigation system and an air data system) and a differential GPS, with centimetric precision, denoted as Precision Positioning System (PPS) and needed for accurate relative navigation during take-off and landing operations.

Each measurement (M) is computed starting from its simulated true value ($\overline{M}$) taken from the models of the helicopter, the ship or the atmosphere.

For what concerns the inertial navigation sensor, each true variable to be measured is filtered and sampled. Then it is corrupted by introducing a scale factor deviation (C_{SF}), a bias (e_{bias}), white noise (e_{white}) and an additive magnetic declination error (e_{dec}), which is zero for all the measurements but the helicopter heading:

$$M = C_{SF}\overline{M} + e_{bias} + e_{white} + e_{dec} \quad (20)$$

The air data measurements are generated through the relation:

$$M = \overline{M} + e_{bias} + e_{white} \quad (21)$$

Concerning the GPS sensor, it is simulated corrupting the true measurements with bias, white noise and diluition of precision error (e_{DOP}):

$$M = \overline{M} + e_{bias} + e_{white} + e_{DOP} \quad (22)$$

All the additive errors in (20), (21) and (22) are stochastic and derived from the specification data sheet of the real sensor.

In the GPS model, these errors depend on the configuration of the sensor, which can work in SPS (Standard Positioning Service), DGPS (Differential GPS) and RTK (Real Time Kinematic) mode. The model also allows injecting a failure which degrades the precision of the sensor from RTK mode (also denoted as Precision Positioning System) to SPS mode.

3 Code Generation and Verification

As said, some of the developed models (e.g. ship, ship air wake and GPS sensor) were automatically software coded after the implementation in Matlab/Simulink, in order to allow their integration into the Finmeccanica real-time Software-In-the-Loop (SIL) simulator, which is used to test on ground the GNC prototype. Figure 3 shows the applied code generation process and testing methodology.

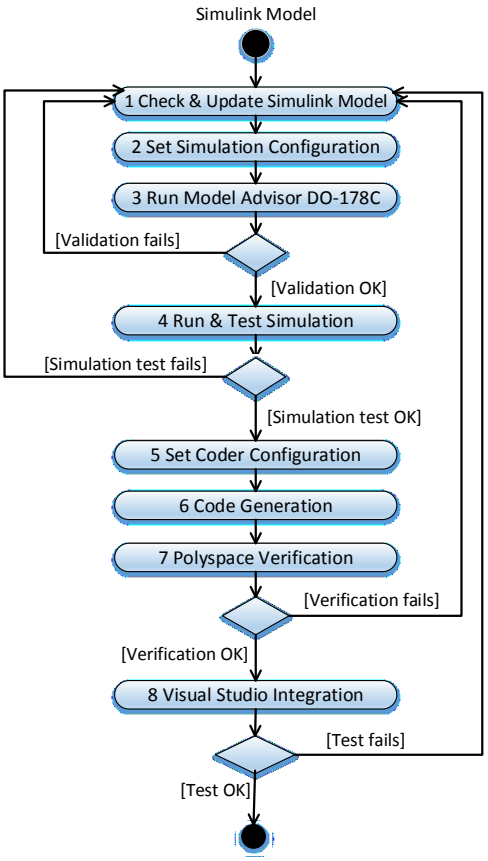


Figure 3. Code generation operational flow.

The flow starts with the selection of the Simulink model from which the code shall be generated. This model shall follow Finmeccanica proprietary design rules and specifications; to this end, a proprietary Simulink library have been developed and used to implement the models.

In step 1, the Model Update command in Simulink environment allows to check for errors and warnings. Then, the configuration settings are applied by running a Matlab script (step 2), that is customized to make the Simulink model compliant to the DO-178C standard. This compliance is verified in step 3 by means of the Mathworks Model Advisor tool. Next, the unit test for each Simulink model is performed, still in Simulink environment (step 4). In step 5, a proprietary Matlab script defines the Code Configuration settings; then the source C Code of the model is automatically generated (step 6) using Real Time Workshop. The Mathworks Polyspace tool is applied in step 7, to perform a static analysis of the generated code in order to check the absence of overflow, divide by zero, out of bounds array access, and other kind of run-time errors. If the generated code passes Polyspace tests, it can be integrated into Microsoft Visual Studio Environment

(step 8) to be tested with the same test vectors used in Step 4. Finally, the test outputs of step 4 and step 8 are compared, in order to check the correctness of the generated code.

After that, the model code can be integrated into the final detailed simulation model, being sure that it performs exactly as the simulation environment used for design.

4 Simulation Results

The principal phenomena that influenced the design of the trajectories and tuning of the tracking algorithm for automatic take-off and landing are the wake phenomenon near the ship, the PPS availability along the trajectory, the disturbance of the sea waves on the ship deck motion and the performance and dynamic behavior of the helicopter.

The validity of the proposed helicopter model for GNC algorithm design can be demonstrated by Figure 4 where comparison of flight data versus simulation data is reported for attitude.

The differences that can be noted have negligible effects on the algorithm design and preliminary testing, as the helicopter low frequency behavior is almost accurately predicted. Similar results hold for acceleration, not reported here for the sake of brevity.

Moreover, Figure 5 compares the collective deflections in trim condition at 650ft altitude computed by using the model with a validation data set provided by Finmeccanica: the model reproduces quite well the vehicle behavior, confirming the validity of the proposed helicopter thrust model.

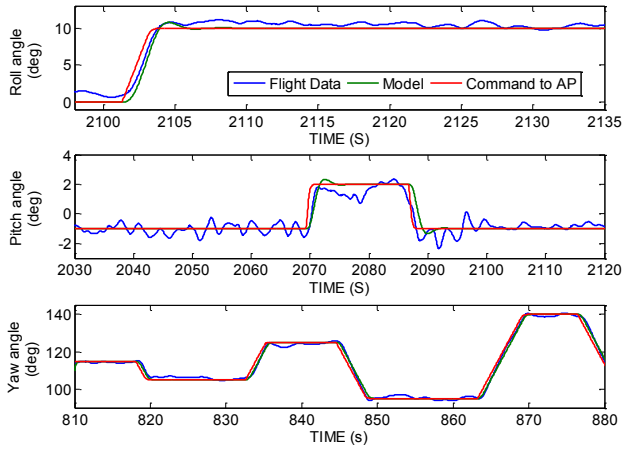


Figure 4. Comparison between simulated and experimental attitudes.

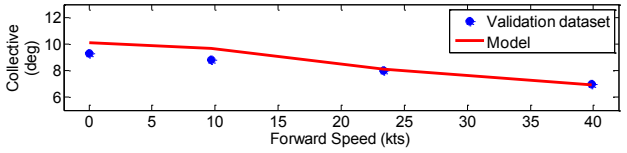


Figure 5. Comparison between simulated collective deflections in trim condition and validation data.

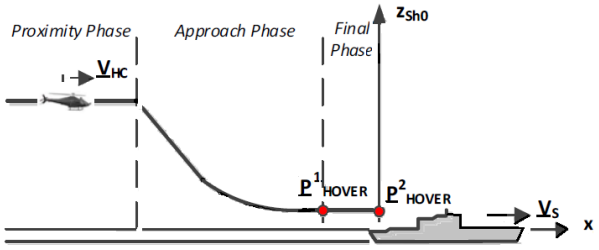


Figure 6. Schematic representation of the landing trajectory.

The other main simulated effects on a sample automatic landing trajectory are also presented below.

The designed landing trajectory, schematically shown in Figure 6, is structured in three phases. In the Proximity phase the helicopter is almost aligned with the ship direction at a desired speed in order to follow properly the descending path to the first relative hovering way point (*Approach phase*). In the Final phase, after the operator acknowledgment, the helicopter moves to the second relative hovering waypoint ($\underline{P}^2_{HOVER}$) and finally lands on the ship deck.

The modelled action of the air wake on the helicopter vertical acceleration, during the automatic landing manoeuvers, is shown in Figure 7.

It is worth to note, in the second graph, how the effect of the air wake is null until the helicopter enters in a proper area (near $\underline{P}^1_{HOVER}$). As said, such area depends on the ship super-structure and wind direction (which in the test is aligned with the ship speed). When the ground operator gives the acknowledge command, the relative distance between the ship and the helicopter decreases while the wake effect increases. The same happens as the relative altitude decreases in the last manoeuver for deck landing.

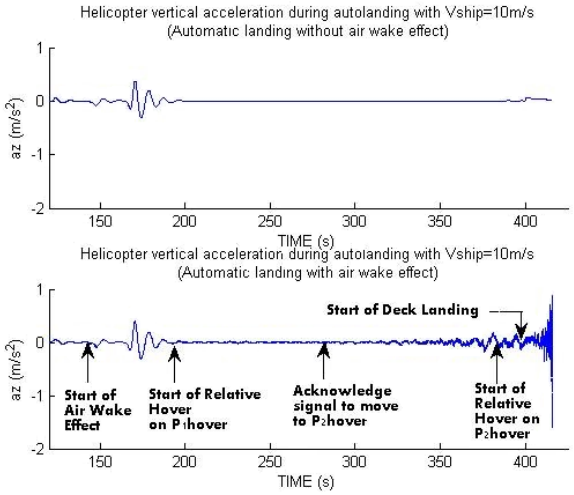


Figure 7. Air wake effect on the helicopter vertical acceleration.

Figure 8 presents the effect of the sea waves on the ship. It refers to Type 23 frigate at two different speeds for see state level six.

As expected, it is highlighted how the frequency and the mean amplitude of ship deck motion decrease with the increase of the ship speed. Indeed wave disturbances are low pass filtered by the ship inertia, and the cut off frequency of this filter decreases when ship speed increases. Such movements are taken into account in the last part of the landing manoeuver when the helicopter waits for a quiescent state of the ship deck before landing.

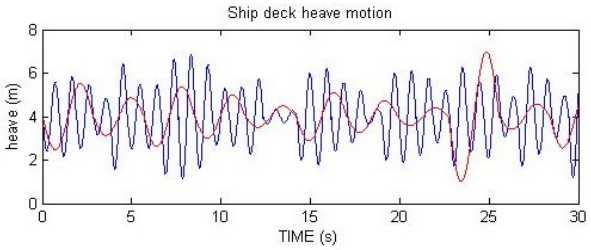


Figure 8. Ship deck motion under sea state level six at 30 knots ship speeds (red) and null ship speed (blue).

5 Conclusions

Accurate, yet simple, simulation environments are fundamental tools to develop GNC algorithms.

This paper presented an effective simulation environment to be used specifically for design, preliminary test and software implementation of automatic take-off and landing algorithms on a ship deck.

With reference to rotary-wing applications, an original modelling approach based on both empirical relations and appropriate mathematical formulation has been proposed that still accurately reproduce helicopter and ship transactional motion and ship air wake.

Simulation results demonstrate effectiveness and accuracy of such modelling approach.

The developed simulation environment contributed to reduce design time, risks and costs of automatic take-off and landing algorithms on a ship deck that were successfully tested in flight.

Future work will be focused on the refinement of the model's parameters (especially for what concern the ship air wake model) based on the analysis of flight data.

References

G. D. Carico, R. Fang, R. S. Finch, W. P. Geyer Jr., H. W. Krijns, and K. R. Long. Helicopter/Ship Qualification Testing, *RTO AGARDograph 300, Flight Test Techniques Series*, 22, 2003.

V. Gavrilets. *Autonomous Acrobatic Maneuvering of Miniature Helicopters*, PhD thesis, MIT, 2003.

L. H. Holthuijsen. *Waves in oceanic and coastal waters*, Cambridge University Press, 2007. doi:10.1017/CBO9780511618536.

W. Johnson. *Helicopter Theory*, Dover, New York, 1994.

C. H. Kääriä. *Investigating the Impact of Ship Superstructure Aerodynamics on Maritime Helicopter Operations*, PhD thesis, University of Liverpool, 2012.

T. von Kármán. Progress in the Statistical Theory of Turbulence, *Proceedings of the National Academy of Sciences*, 34 (11): 530-539, 1948. doi:10.1073/pnas.34.11.530.

D. Lee. *Simulation and Control of a Helicopter Operating in a Ship Airwake*, PhD thesis, Pennsylvania State University, 2005.

Z. Li. *Path Following with Roll Constraints for Marine Surface Vessels in Wave Fields*, PhD thesis, University of Michigan, 2009.

B. W. McCormick. *Aerodynamics, Aeronautics, and Flight Mechanics*, John Wiley & Sons, New York, 1995.

MIL-F-8785C. *Military Specification Flying Qualities for Piloted Airplanes*, 1991.

G. D. Padfield. The making of helicopter flying qualities: A requirements perspective, *The Aeronautical Journal*, 102 (1018): 409 – 437, 1998.

G. D. Padfield. *Helicopter Flight Dynamics: The Theory and Application of Flying Qualities and Simulation Modeling*, AIAA Education Series, Virginia, 1996.

T. Perez. *Ship Motion Control*, Springer, 2005. doi:10.1007/1-84628-157-1.

M. B. Tishler and R. K. Remple. *Aircraft and Rotorcraft System Identification*, AIAA Education Series, Virginia, 2006. doi:10.2514/4.868207.

X. Yang, H. R. Pota, and M. Garratt. Design of a Gust-Attenuation Controller for Landing Operations of Unmanned Autonomous Helicopters, *In Proceedings of 18th IEEE International Conference on Control Applications*, Saint Petersburg, Russia, 2009. doi:10.1109/CCA.2009.5281074.

Simulation Model of a Piston Type Hydro-Pneumatic Accumulator

Juho Alatalo Toni Liedes Mika Pylvänäinen

Mechatronics and Machine Diagnostics, University of Oulu, Finland,
{juho.alatalo,toni.liedes,mika.pylvanainen}@oulu.fi

Abstract

Hydro-pneumatic accumulators are used to improve the features of different kinds of hydraulic systems and they are common in industry and mobile applications. In order to include functionality of accumulators to hydraulic system models, an accurate yet light simulation model of hydro-pneumatic accumulator is needed. In this paper a simulation model of a piston type hydro-pneumatic accumulator is presented. The simulation model takes into account of the behavior of friction, nitrogen gas and hydraulic fluid. The simulation model was validated by comparing the simulation results to measurement results obtained from laboratory tests, and strong correlation was found between them. The model is suitable for researchers as well as for engineers in designing work in industry.

Keywords: gas-charged accumulator, computer simulations, modelling

1 Hydro-pneumatic Accumulator

Storing the energy in hydraulic fluid is very tricky. That is why there is a separate component for storing and releasing energy in hydraulic systems, an accumulator. In the accumulator energy can be bound for example to the gas pressure or potential energy of the spring. This study focuses on examining the gas-operated accumulator, so called hydro-pneumatic accumulators.

Hydro-pneumatic accumulators are conventionally utilized in hydraulic systems in several ways. One of the most common way is to improve the features of the hydraulic system, for example, in order to absorb the pressure impulses in the hydraulic line. The other main method of application for the hydro-pneumatic accumulator is to secure the safe operation of the machine, and in applications of this kind the accumulator usually operates as a reserve volume flow source. One of the main advantages of the hydro-pneumatic accumulators is their capability to act as a local and distributed energy storage and volume flow source.

Hydro-pneumatic accumulators are widely used in industry and mobile applications. In industry, the hydro-pneumatic accumulators are common in power plants such as windmills and wave power plants. In mobile applications, hydro-pneumatic accumulators are also common in different kinds of mobile working machines like mine loaders, harvesters and load-haul-dump machines. Hydro-pneumatic suspension is a widespread solution in a variety of mobile applications: cars, trucks, agricultural vehicles and military vehicles.

Three main types of accumulators exist with different kinds of properties: a diaphragm accumulator, a bladder accumulator and a hydro-pneumatic piston accumulator. Bladder and diaphragm accumulators are very much alike, but the piston type accumulator has some pros and cons compared to the two mentioned above. The main disadvantages in a piston type accumulator compared to the bladder and diaphragm type accumulators are its insensitivity to small and rapid pressure fluctuations and the greater weight of the accumulator, which may be essential in mobile machines, but not necessary in an industrial environment. The main advances in a piston type accumulator compared to the bladder and diaphragm type accumulators are its ability to handle a wider range of pressure changes and its ability to be entirely depleted from hydraulic fluid. In addition, a piston type accumulator is less sensitive to temperature changes and its installation orientation is not predefined. Also, using the sensors is easier in piston type accumulators, which means that the state of the accumulator is easier to monitor (Palomäki, 2012). In most applications all of the accumulator types mentioned above can be used, but in research and development the piston type is preferred because of its wider ranges of temperature and pressure, its modifiability, durability and the ability to be easily monitored by sensors. The structure of a piston type hydro-pneumatic accumulator is shown in in Figure 1.

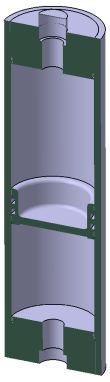


Figure 1. Cross-section of a piston type hydro-pneumatic accumulator.

1.1 Recent Research

The focus of recent studies in this field is on energy consumption and energy storing, and the potential of the hydro-pneumatic accumulator in this research areas has been realized. Even though the hydro-pneumatic accumulator is an old invention, the research and development of piston type accumulators is constantly ongoing and some popular development areas can be introduced: the improvement of piston type accumulator energy storing capacity (Tavares, 2011) and the improvement of a piston type accumulator's efficiency by utilizing heat regeneration (Stroganov and Sheshin, 2011). Both these areas are closely linked with research into energy recovering systems (Achten, 2008; Ancai and Jihai, 2009; Lin et al., 2010; Zhang et al., 2010), which in particular is a contemporary research area in the field of mobile working machines.

Minav et al. (2012) have studied means of improving the energy efficiency of a mobile working machine. In their conclusion the storing of the hydraulic energy in piston type accumulators was seen as a waste of energy in a given type of energy storing method. Instead, Minav et al. proposed that energy should be stored in electric form. Their research shows the need for additional research into hydraulic systems and hydraulic energy storing methods. Instead of abandoning the idea of using hydro-pneumatic accumulators as energy storages in a mobile working machine, it would be economically reasonable to redesign and improve the old accumulator design and inefficient energy storing methods. This is due the fact that hydraulics are still widely favored in industry and especially mobile working machines due to its good power-to-weight-ratio and the ease of maintenance.

1.2 Purpose of This Study

Despite the simplicity of the piston type hydro-pneumatic accumulator's structure, it is fairly complex to model, because from the modelling point of view it is a multi-domain system. This means that it has several different parts: mechanics, thermodynamics, pneumatics and hydraulics.

This study presents an advanced but yet feasible simulation model of a piston type hydro-pneumatic accumulator. The aim of the simulation model is to be an easy-to-use but yet accurate tool for researchers or engineers and designers in industry, for example, providing information on how an accumulator works in larger hydraulic systems, such as energy storing or suspension systems.

2 Logical Structure of Simulation Model

In this section the relevant physical phenomena of a piston type hydro-pneumatic accumulator are examined and the conceptual model is created on which the simulation model is based. The conceptual model is used to ensure that the logical structure of the model is correct and that all the essential elements that describe the behavior of the real hydro-pneumatic accumulator are included in the simulation model. When all the interactions between these elements are properly described, the structure of the simulation model is verified allowing the simulation model to achieve the desired accuracy.

2.1 Physics of Piston Type Hydro-Pneumatic Accumulator

In a piston type hydro-pneumatic accumulator, part of the energy adhered to the hydraulic fluid is converted to the energy adhered to the nitrogen gas with the help of the motion of the piston.

2.1.1 Nitrogen gas

In the regular duty cycle of the hydro-pneumatic accumulator, the volume V_{gas}, pressure p_{gas} and temperature T_{gas} of the nitrogen gas change and their values are dependent on each other. The volume of the gas, naturally, follows the position of the piston and volume and temperature increase and decrease in relation to the position of the piston. The nitrogen gas interacts with its environment not only via piston displacement, but also by the transition of heat energy through the piston wall. The heat exchange between the environment and the gas depend on the temperatures of the gas and the environment. In this case the environment of the gas is considered to be the interior of the cylinder wall. Also the amount of nitrogen gas usually decreases slowly over the time, due to the leakage past the piston or through the gas valve.

2.1.2 Mechanical Structure

There are two main mechanical structures in piston type hydro-pneumatic accumulators: the piston and cylinder wall. The piston is the only moving part of the mechanical structure and the kinetic energy is bound to the piston when it is moving. There is friction between the seal of the piston and the cylinder wall, which should be noted especially when the piston is starting to move from the rest position. The friction also heats up the piston and cylinder wall. The seal of the piston is a flexible element and together with piston inertia and static friction it causes the piston type accumulator to be insensitive to the small and high frequency fluctuations of pressure. The cylinder wall mainly acts as a thermal energy reservoir, which interacts with nitrogen gas and the ambience of the accumulator. All the mechanical parts are subject to constant changes in pressure and temperature, which affects the dimensions of these mechanical parts.

2.1.3 Hydraulic Fluid

The pressure, volume and the temperature of the hydraulic fluid also change during the regular duty cycle of the accumulator, but compared to nitrogen gas the changes in volume and temperature are much less significant. If the hydraulic fluid contains a lot of air bubbles, changes of volume occur at the lower pressure levels. There is also

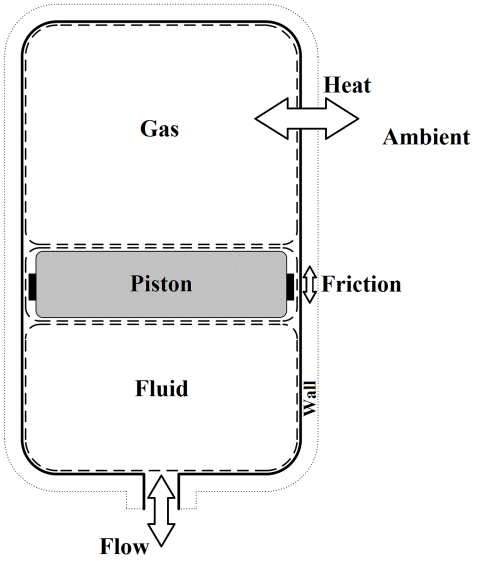


Figure 2. Conceptual model of a hydro-pneumatic accumulator.

one major difference between the nitrogen gas and the system that the hydraulic fluid forms: the amount of hydraulic fluid in a hydro-pneumatic accumulator varies dramatically during the regular duty cycle of the hydro-pneumatic accumulator due to the input and output flow. When flowing, the hydraulic fluid has some kinetic energy adhered to itself and also some dissipation occurs.

2.2 Conceptual Model

In this study the accumulator is considered to be a system that consists of several subsystems. These subsystems represent the essential elements needed to create the model. Figure 2 shows the subsystems that form the whole system of an accumulator. The subsystems are nitrogen gas, piston and hydraulic fluid. Everything outside of these subsystems is assumed to be ambient. It is important to note that some assumptions are made in order to simplify the model. It is good to keep in mind that we want the model to be feasible but yet accurate enough, so some assumptions and idealizations have to be made. One main simplification is that cylinder walls are not part of this model. The walls of the cylinder store thermal energy and play quite a significant role in the accumulator. The role of the cylinder wall can be taken into account when choosing the mathematical representation of the nitrogen gas, which will be discussed in the next section. All the material in each subsystem is predicted to be homogenous, which means that air bubbles in the hydraulic fluid are not part of the model. Changes in the dimensions of the cylinder wall and piston are not taken into account. Also the subsystem of gas is assumed to be a closed system, which means that the amount of substance in this subsystem is constant. The kinetic energy adhered to the motion of the hydraulic fluid is considered negligible.

3 Mathematical Model

This simulation model was made using MATLAB/Simulink, which is advisable for modelling a multi-domain system of this kind. The approach of this simulation model was chosen to be a lumped model, which at best keeps the simulation model light. In this section we will go through the equations used in this model.

3.1 Equations for Nitrogen Gas

In this simulation model the nitrogen gas was described with a real gas model, which has been used in several investigations. Heat transfer in a hydro-pneumatic accumulator has been studied by Els and Grobbelaar (1999), Pourmovahed and Otis (1990) and Giliomee (2003). The advantages of this real gas model compared to the commonly used ideal gas model in a piston type hydro-pneumatic accumulator have been investigated by Puddu and Paderi (2013) and Kroneld (2018) used real gas model in his research.

In this real gas model, the pressure, volume and the temperature of the gas are described using three equations. (1) describes the heat transfer between the gas and the environment,

$$\dot{T}_{\text{gas}} = \frac{(T_{\text{amb}} - T_{\text{gas}})}{\tau_{\text{therm}}} - \frac{\dot{v}}{C_v}\left[\frac{RT_{\text{gas}}}{v}\left(1+\frac{b}{v^2}\right) + \frac{1}{v^2}\left(B_0 RT_{\text{gas}} + \frac{2C_0}{T_{\text{gas}}^2}\right) - \frac{2c}{v^3 T_{\text{gas}}^2}\left(1+\frac{\gamma}{v^2}\right)e^{-\frac{\gamma}{v^2}}\right], \quad (1)$$

where $\dot{T}_{\text{gas}}$ is change in gas temperature, T_{amb} is ambient temperature, T_{gas} is gas temperature, τ_{therm} is thermal time constant, $\dot{v}$ is change in specific volume of the gas, C_v is constant volume-specific heat capacity of gas, v is the specific volume of the gas and parameters R, b, B_0, C_0, c, and γ are specific constants for nitrogen gas and their values can be found in (Giliomee, 2003).

(2) is the Benedict-Webb-Rubin -equation of state (Cooper and Goldfrank, 1967),

$$p_{\text{gas}} = \frac{RT_{\text{gas}}}{v} + \left(\frac{B_0 RT_{\text{gas}} - A_0 - \frac{C_0}{T_{\text{gas}}^2}}{v^2}\right) + \left(\frac{bRT_{\text{gas}} - a}{v^3}\right) + \frac{a\alpha}{v^6} + \left[\frac{c\left(1+\frac{\gamma}{v^2}\right)e^{\frac{-\gamma}{v^2}}}{v^3 T_{\text{gas}}^2}\right], \quad (2)$$

where p_{gas} is gas pressure and parameters A_0, a and α are specific constants for nitrogen gas and their values can be found in (Giliomee, 2003).

(3) is the equation for the specific heat capacity of the

gas,

$$C_v = R\left[\frac{N_1}{T_{\text{gas}}^3}+\frac{N_2}{T_{\text{gas}}^2}+\frac{N_3}{T_{\text{gas}}}+(1-N_4)+N_5T_{\text{gas}} + N_6T_{\text{gas}}^2+N_7T_{\text{gas}}^3+\frac{N_8\left(\frac{N_9}{T_{\text{gas}}}\right)^2 e^{\frac{N_9}{T_{\text{gas}}}}}{\left(\frac{N_9}{e^{T_{\text{gas}}}-1}\right)^2}\right], \quad (3)$$

where parameters N_1-N_9 are specific constants for nitrogen gas and their values can be found in (Giliomee, 2003).

As mentioned before, the cylinder wall is not included in this model. The lack of this kind thermal energy storage can be taken into account with the thermal time constant, which must be determined by measuring. This restricts the range of use of the model to situations where the system has been settled to the thermal balance e.g. for sinusoidal use. Also the ambient temperature must be measured. The mass of the nitrogen gas must also be measured in order to calculate the specific volume of the gas.

The way how the state variables of the real gas model are connected and how they depend on the other state variables of the simulation model can be seen from Figure 3.

3.2 Equations for Piston

The piston in this simulation model is constantly subjected to three different forces: the force caused by the hydraulic fluid pressure F_{hyd}, the force caused by the gas pressure F_{gas} and friction force F_μ. This means that the total force on piston F_{piston} can be written as:

$$\begin{aligned} F_{\text{piston}} &= F_{\text{hyd}} - F_{\text{gas}} \pm F_\mu \\ &= \Delta p_{\text{piston}} A_{\text{piston}} \pm F_\mu \\ &= (p_{\text{hyd}} - p_{\text{gas}}) A_{\text{piston}} \pm F_\mu, \end{aligned} \quad (4)$$

where Δp_{piston} is pressure differentiation over the piston, A_{piston} is the area of the piston, p_{hyd} is pressure of the hydraulic fluid and p_{gas} is pressure of the gas.

From Newton's second law we get:

$$F_{\text{piston}} = m_{\text{piston}}\ddot{x}_{\text{piston}}, \quad (5)$$

where m_{piston} is the mass of piston and $\ddot{x}_{\text{piston}}$ is the acceleration of piston.

The mass of the piston is easy to measure and no other parameters are needed to determine the piston.

The way the state of the piston depends on the other state variables of the simulation model of the accumulator can be seen from Figure 3.

3.3 Equations for Friction

The friction in this simulation model has been described with the friction model introduced by Olsson (1996). The friction model for the pre-sliding displacement is described with the help of bristles. After the bristles have deflected a certain amount, the surfaces start to slide with respect to each other. Friction plays important role in a piston type accumulator especially when the piston is changing direction or is at rest, and without such a complex friction model the accuracy of the simulation model would not be satisfying.

The friction force F_μ is written as:

$$F_\mu = \sigma_0 z + \sigma_1(v_{\text{frict}})\dot{z} + F_{\text{v}} v_{\text{frict}}, \quad (6)$$

where σ_0 is the stiffness of the bristle, z is the deflection of the bristle, $\sigma_1(v_{\text{frict}})$ is velocity dependent damping coefficient, $\dot{z}$ is the rate of change of deflection of the bristle, F_{v} is viscous friction coefficient and v_{frict} is relative velocity between surfaces.

The velocity dependent damping coefficient can be written as:

$$\sigma_1(v_{\text{frict}}) = \sigma_1 e^{-\left(\frac{v_{\text{frict}}}{v_{\text{d}}}\right)^2}, \quad (7)$$

where σ_1 is damping coefficient and v_{d} is the velocity when the damping affects.

The damping coefficient σ_1 can be chosen by using the equation:

$$0.5 \ldots 2\sqrt{\sigma_0 m_{\text{piston}}}. \quad (8)$$

The rate of change of deflection of the bristle can be written as:

$$\dot{z} = v_{\text{frict}} - \frac{|v_{\text{frict}}|}{g(v_{\text{frict}})} z, \quad (9)$$

where the equation $g(v_{\text{frict}})$ can be written as:

$$g(v_{\text{frict}}) = \frac{1}{\sigma_0}\left[F_{\text{C}} + (F_{\text{S}} - F_{\text{C}}) e^{-\left(\frac{v_{\text{frict}}}{v_{\text{S}}}\right)^2}\right], \quad (10)$$

where F_{C} is Coulomb friction, F_{S} is static friction and v_{S} is Stribeck velocity.

The parameters in the friction model are strongly dependent on the materials used and they must be figured out by measuring in order to make the friction model work. These measured parameters are F_{C}, F_{S}, F_{v}, σ_0, σ_1, v_{d} and v_{S}. The way the state variables of the friction model are connected to each other and how they depend on the other state variables of the simulation model of the accumulator can be seen from Figure 3.

3.4 Equations for Hydraulic Fluid

To figure out the pressure of the hydraulic fluid, the following equation were used (Merritt, 1967):

$$\dot{p}_{\text{hyd}} = \frac{B_{\text{hyd}}}{V_{\text{hyd}}}\left(\sum q - \dot{V}_{\text{hyd}}\right), \quad (11)$$

where $\dot{p}_{\text{hyd}}$ is the rate of change of the pressure of the hydraulic fluid, B_{hyd} is the bulk modulus of the hydraulic fluid, q is flow and $\dot{V}_{\text{hyd}}$ is the rate of change of the volume of the hydraulic fluid.

The pressure of the hydraulic fluid can be solved by integrating (11). The bulk modulus must be defined before this model can be used as a part of the simulation model

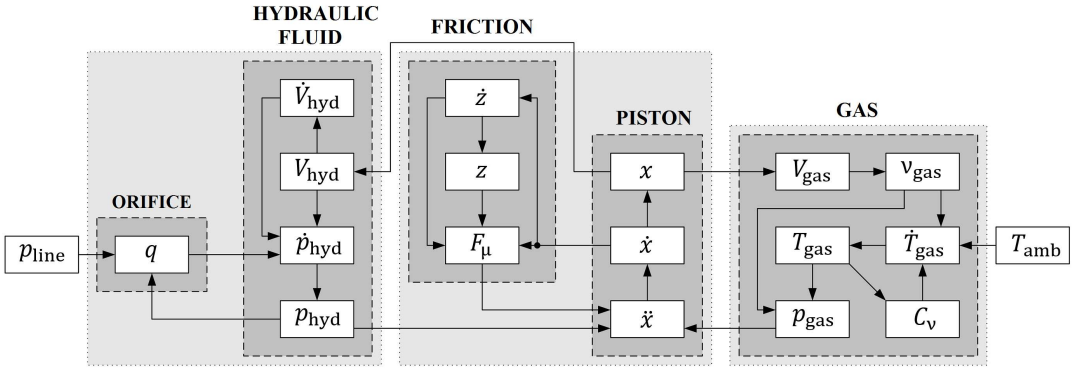


Figure 3. State variables and their connections in the simulation model.

and it is assumed to be constant. The way the state variables of the hydraulic fluid model are connected to each other and how they depend on the other state variables of the simulation model of the accumulator can be seen from Figure 3.

3.5 Equations for Orifice

If it is desired to use the pressure of the hydraulic line as an input for the whole simulation model, the input information is needed in order to convert to the quantity of the flow q before we can utilize it in a hydraulic fluid model. For this purpose the orifice model is placed between input pressure information and hydraulic fluid model. Also this model can be used to describe the dissipation that occurs when hydraulic fluid is flowing.

The orifice model used was introduced by Åman (2011) and it is a two regional orifice model that describes the laminar flow more accurately than the commonly used turbulent flow model. The difference between these two models becomes notable especially at zero flow and when the flow is changing direction. The model describes the flow with a piecewise defined equation:

$$q = \begin{cases} a_1\Delta p_{\text{orif}} + a_2\Delta p_{\text{orif}}^2 + a_3\Delta p_{\text{orif}}^3 & \text{if } |\Delta p_{\text{orif}}| < |\Delta p_{\text{trans}}| \\ K\sqrt{|\Delta p_{\text{orif}}|}\,sign(\Delta p_{\text{orif}}) & \text{if } |\Delta p_{\text{orif}}| \geq |\Delta p_{\text{trans}}| \end{cases}, \tag{12}$$

where Δp_{orif} is the pressure difference over the orifice.

When the density of the hydraulic fluid is assumed to be constant, the parameter K can be written as:

$$K = C_{\text{q}} A_{\text{orif}} \sqrt{\frac{2}{\rho_{\text{hyd}}}}, \tag{13}$$

where C_{q} is discharge coefficient, A_{orif} is the area of the orifice and ρ_{hyd} is the density of the hydraulic fluid.

According to Åman, the pressure limit Δp_{trans} can be defined as:

$$\Delta p_{\text{trans}} = \frac{Re^2 \upsilon^2 \pi \sqrt{\rho_{\text{hyd}}}}{\sqrt{32} C_{\text{q}} K}, \tag{14}$$

where Re is the Reynolds number and υ is kinematic viscosity of the hydraulic fluid.

The parameters a_1, a_2 and a_3 can be expressed as matrix $\mathbf{A}$:

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \\ &= \begin{bmatrix} \frac{5K}{4\sqrt{\Delta p_{\text{trans}}}} & 0 & -\frac{K}{4\Delta p_{\text{trans}}^{\frac{5}{2}}} \end{bmatrix}. \end{aligned} \tag{15}$$

The parameters C_{q}, Re, ρ_{hyd} and υ must be defined before this model can be used as part of the simulation model. Also the parameter A_{orif} must be adjusted so that the pressure drop over the orifice corresponds to the measured values. The way how the variable of the orifice model depends on the other variables of the simulation model of the accumulator is shown in Figure 3.

All the variables presented in Figure 3 can be solved either by means of the given equations or by integrating them. The inputs for the whole system are the pressure of the hydraulic line and environment temperature. If necessary, the flow can be used as an input too, but then the pressure of the hydraulic line and the orifice model must be ignored.

4 Modelling in MATLAB/Simulink

The layout of the simulation model built in MATLAB/Simulink follows the structure represented in Figure 3. In order to keep the layout of the simulation model lucid, the subsystems were used in Simulink. The named blocks in Figure 3 (Orifice, Hydraulic fluid, Friction, Piston and Gas) represent the contents of these subsystems. As an example of the method how the equations in the orifice subsystem are used in Simulink is shown in Figure 4. As can be seen from Figure 4 the inputs for the subsystem are the pressure of the hydraulic line, p_{line}, and the pressure of the hydraulic fluid, p_{hyd}. The output of this subsystem is the flow q. The parameters needed in the subsystem are read from a separate file, and the new set of parameters are extremely easy to feed to the model by just modifying the file where the parameters are read from. This makes the use of this simulation model flexible. The rest of the Simulink model is built by using the same kind of approach for every subsystems as described above.

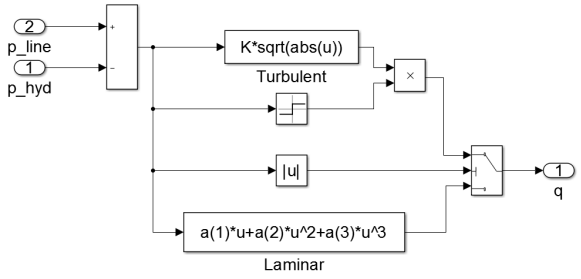


Figure 4. The orifice subsystem in Simulink.

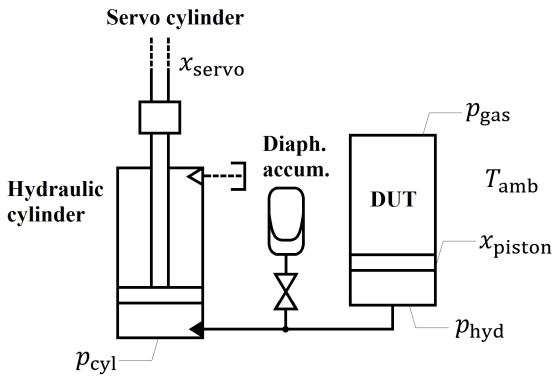


Figure 5. Schematic figure of the testing setup.

5 Calibration and Validation of the Simulation Model

The equations used in the simulation model were introduced in the previous section. These equations contained unknown parameters, excluding some parameters in the real gas model, which had to be defined. In order to define these parameters for this application, a testing setup was built and measurements were performed. The goal is to determine these parameters so accurately that our accumulator model is well calibrated and its performance can be validated by comparing the simulation results to the result obtained from the measurements for the real piston type hydro-pneumatic accumulators. The parameters were determined for five different piston type hydro-pneumatic accumulators, with piston diameters 60 mm, 80 mm, 100 mm, 125 mm and 140 mm and a nominal volume of 2 liters.

5.1 Testing Setup

The testing setup was built in the testing laboratory of the department of mechanical engineering at the University of Oulu. The schematic picture of the testing setup can be seen in Figure 5. The testing setup is a closed hydraulic system, which consist of a hydraulic cylinder that is directly connected to the piston type hydro-pneumatic accumulator. There is also a parallel diaphragm hydro-pneumatic accumulator that is only used in certain measurement runs. The piston of the hydraulic cylinder is driven by a hydraulic servo.

The pressure of the hydraulic fluid was measured directly from the bottom of the hydraulic cylinder and the piston type hydro-pneumatic accumulator and the pressure of the nitrogen gas was measured from the gas side of the piston type hydro-pneumatic accumulator. The pressures were measured using IFM PT9541 sensors. The position of the piston of the hydro-pneumatic accumulator was measured with a non-contact sensor in order not to disturb the performance of the accumulator with additional friction. The position of the piston was measured with a Temposonics MH Series sensor. Also the position of the servo cylinder and the outside temperature were measured. The servo position sensor was HBM 1-WA/500MM-L and the temperature was measured with an ordinary K-type thermocouple. The wall of the 140 mm diameter accumulator was so thick that the Temposonics sensor was unable to measure the position of the piston. Instead, the position of the piston of the 140 mm diameter accumulator was calculated from the position of the servo cylinder and by estimating the compression of the hydraulic fluid. The data acquisition was performed using NI cDAQ-9174, with NI 9215 and NI 9211 modules and with 250 Hz sampling rate. The variables at Figure 5 represent the measured quantities.

5.2 Laboratory Tests

Three different kinds of calibration runs were performed with the testing setup in order to determine the parameters named in Section 3.

The parameters F_C, F_S, F_v, v_d and v_S for the friction model were determined with the sinusoidal motion input excitation of the servo cylinder. The Stribeck curve could be created from the information on pressure differential over the hydro-pneumatic piston and the position information of the piston. The values for the parameters could be derived from the Stribeck curve.

The parameters σ_0 and σ_1 for the friction model were determined with the help of a diaphragm hydro-pneumatic accumulator. The pressure in the hydraulic system was raised slowly and the pressure and the maximum presliding deflection of the piston seals were measured. From this information the desired parameters could be derived.

To determine the thermal time constant τ_{therm} of the piston type hydro-pneumatic accumulator, step-type excitation was introduced to the accumulator. The pressure and the volume of the nitrogen gas was measured and the value of the τ_{therm} in the simulation model was iterated in so that the pressure of the simulation model behaved accurately enough.

5.3 Validation

The simulation models were validated by comparing their results to the measured values. Measured and simulated results are presented Figure 6 – Figure 9 and as can be seen, the simulation results correspond really well to the

measured data. Some of the curves appear to be one on the other, but on the *pV* chart some difference can be seen. Note that in Figure 8 the position of piston is estimated.

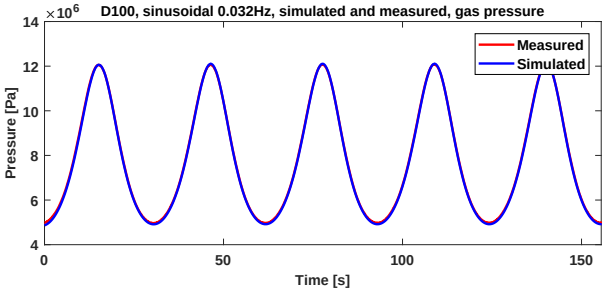


Figure 6. The gas pressures of the 100 mm diameter accumulator when the input for the accumulator is the sinusoidal 0.032 Hz flow.

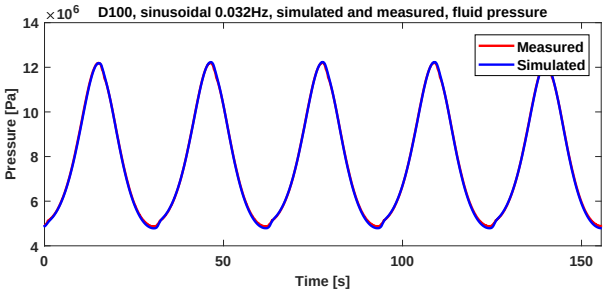


Figure 7. The hydraulic fluid pressures of the 100 mm diameter accumulator when the input for the accumulator is the sinusoidal 0.032 Hz flow

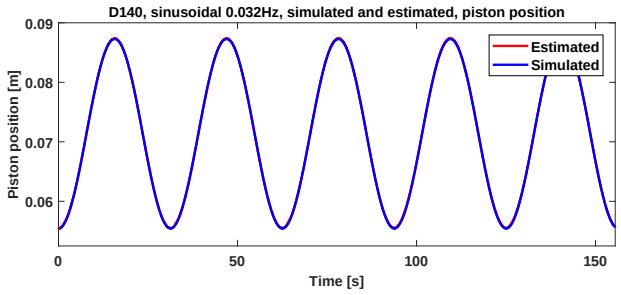


Figure 8. The piston position of the 140 mm diameter accumulator when the input for the accumulator is the sinusoidal 0.032 Hz flow.

6 Conclusions

The simulation model of the piston type hydro-pneumatic accumulator met the demand for an easy-to-use but yet accurate simulation model and it can be used either in research or in mechanical design work. The results of the simulation model were very accurate and there was strong correlation between the simulation and the measured data. The model itself appeared to be fairly easy to simulate. The simulation model is capable of simulating the performance of the piston type hydro-pneumatic accumulator in situations where the system is in thermal balance.

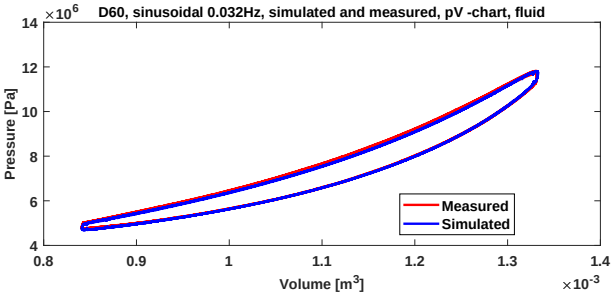


Figure 9. The *pV* chart of the 60 mm diameter accumulator when the input for the accumulator is the sinusoidal 0.032 Hz flow.

Real-time model-based condition monitoring can be regarded as an important method in future and these kinds of simulation models of different components play an important role in it. This simulation model provides a good foundation when heading towards this new practice.

There is need for further development in modelling the cylinder walls, which probably also allows the possibility to simulate situations where the accumulator is on the transition between thermal balance situations, thus extending the range of use of the simulation model.

References

P. A. J. Achten. A serial hydraulic hybrid drive train for off-road vehicles. In *Proceeding s of the National Conference on Fluid Power*, volume 51, pages 515 – 521, 2008.

Y. Ancai and J. Jihai. Research on the regenerative braking control strategy for secondary regulation hydrostatic transmission excavators. In *Proceedings of the 2009 IEEE International Conference on Mechatronics and Automation*, pages 4600 – 4604, 2009.

H. W. Cooper and J. C. Goldfrank. B-W-R constants and new correlations. *Hydrocarbon Processing*, 46(12):141 – 146, 1967.

P. S. Els and B. Grobbelaar. Heat transfer effects on hydropneumatic suspension systems. *Journal of Terramechanics*, 36: 197 – 205, 1999.

C. L. Giliomee. *Analysis of a Four State Switchable Hydro-Pneumatic Spring and Damper System*. PhD thesis, University of Pretoria, 2003.

P. Kroneld. *Energy Managment in Active Suspension Systems*. PhD thesis, University of Oulu, 2018. [unpublished].

T. Lin, Q. Wan, B. Hu, and W. Gong. Research on the energy regeneration systems for hybrid hydraulic excavators. *Automation in Construction*, 19:1016 – 1026, 2010.

H. E. Merritt. *Hydraulic control systems*. New York : Wiley, 1967. ISBN-13:978-0471596172.

T. A. Minav, A. Virtanen, L. Laurila, and J. Pyrhönen. Storage of energy recovered from an industrial forklift. *Automation in Construction*, 22:506 – 515, 2012.

H. Olsson. *Control Systems with Friction*. PhD thesis, Lund Institute of Technology, 1996.

H. Palomäki. Paineakkujen testilaitteiston suunnittelu. Master's thesis, Tampere University of Technology, 2012. [in Finnish].

A. Pourmovahed and D. R. Otis. An experimental thermal time-constant correlation for hydraulic accumulators. *Journal of Dynamic Systems, Measurement, and Control*, 112(1):116 – 121, 1990.

Pi. Puddu and M. Paderi. Hydro-pneumatic accumulators for vehicles kinetic energy storage: Influence of gas compressibility and thermal losses on storage capability. *Energy*, 57: 326 – 335, 2013.

R. Åman. *Methods and Models for Accelerating Dynamic Simulation of Fluid Power Circuits*. PhD thesis, Lappeenranta University of Technology, 2011.

A. Stroganov and L. Sheshin. Improvement of heat-regenerative hydraulic accumulators. *Ventil*, 17(4):322 – 332, 2011.

F. T. Tavares. *Thermally Boosted Concept for Improved Energy Storage Capacity of a Hydro-Pneumatic Accumulator*. PhD thesis, University of Michigan, 2011.

X. Zhang, S. Liu, Z. Huang, and L. Chen. Research on the system of boom potential recovery in hydraulic excavator. In *Proceedings of 2010 International Conference on Digital Manufacturing & Automation*, volume 2, pages 3030 – 306, 2010.

Controlling Emergency Vehicles in Urban Traffic with Genetic Algorithms

Monica Patrascu[1] Vlad Constantinescu[1,2] Andreea Ion[1]
[1]Department of Automatic Control and Systems Engineering, University Politehnica of Bucharest, Romania
[2]Institute of Space Science, Bucharest, Romania
{monica.patrascu, vlad.constantinescu, andreea.ion}@acse.pub.ro

Abstract

Emergency officers could often benefit from a route planning system that is based on constant traffic monitoring and complex decision making, seeking to give victims another breath of hope by assisting emergency units with reaching them on time. The main challenge is providing responses in a continuously evolving environment within a prescribed time frame, while using limited resources and information that is often incomplete or uncertain. This paper presents a route control concept for emergency vehicles through urban traffic. The proposed genetic controller is designed to dynamically reassess the route while the vehicle passes through the road network, continuously generating new routes based on current traffic. The algorithm is tested in an agent based simulation model that includes both traffic participants and a distributed traffic control system.

Keywords: genetic algorithm, emergency response, control systems, distributed control, agent based simulation model

1 Introduction

In complex and distributed urban environments, the services that provide quality of life and safety have to deal with unpredictable events and incomplete data. Moreover, intelligent transport systems are becoming increasingly important as they aim to provide solutions to crucial issues related to transportation networks, such as congestion and various incidents. One of the most important activities in the protection of human life is the intervention of emergency responders, for which an important issue in the unpredictable urban road networks is the time required for an emergency vehicle to reach an event scene. Congestion and the various obstacles that may appear during the journey on the chosen path can increase travel time and therefore reduce the chances of ensuring the safety of human life (Blackwell et al, 2002; Pons et al 2005; Sladjana et al 2011; Rushworth et al 2014).

Thus, re-calculating the routes of emergency vehicle during their journey based on environmental changes is a way to avoid these obstacles.

Real-time decision problems are also playing an increasingly important role in transportation management, as advances in communication and information technologies allow real-time information to be quickly obtained and processed. Therefore, dynamic vehicle route generation has become more and more efficient, especially in urban areas.

The problem of finding the most efficient routes for the quick access of the emergency vehicles in the current urban traffic is very important in terms of protecting and saving human lives. From an economic and social point of view, implementing the developed algorithms would increase the number of saved lives, reduce congestion and accident risk, would reduce fuel consumption and the time spent in traffic and by doing so, would also reduce the number of people affected by stress on the road.

The most important technological benefit regarding evolutionary computing is the possibility to integrate techniques typically associated with modeling complex systems in representing the possible solutions to optimization problems solved with the help of evolutionary algorithms. This opens the way to using these class of algorithms for solving problems that cannot be modeled using formal techniques and that can only be solved by using heuristic methods. The future applications of evolutionary computing are not restricted to vehicle routing; they include different other optimization problems, from designing control systems for processes affected by non-linearity and uncertainties, modeling complex and biological processes, algorithms for the optimization of sensor spreading over an area, to designing and tuning the command rules for distributed control systems applied to large-scale processes.

The initial route generation problem has been initially regarded as a variant of the travelling salesman problem (Dantzing et al 1959). Beside this classical formulation of the routing problem, a series of other approaches have been studied (Toth et al 2002). For route reconfiguration, the initial studies (Seguin et al 1997) have first taken into account the static routing problem (Psaraftis 1980; Madsen et al 1995), followed by more in depth analyses of the differences between dynamic and static routing (Psaraftis, 1988; Goel et al 2006).

Another perspective for solving the problem of dynamic vehicle routing takes into account evolutionary computing algorithms, either by using algorithms inspired from biology (Potvin, 2009) or

machine learning techniques such as supervised learning or genetic programming (Benyahia et al 1998). A technique based on hybrid genetic algorithms was used (Jih et al 1999) for solving the problem of routing a vehicle with size constrains. Another example relies on Dijkstra algorithms (Barrachina et al 2014) and evolutionary strategies for finding an optimum path in a short period of time. Another approach presented in the project Emergency Vehicle Priority implies controlling the traffic lights in favor of the emergency crews (White, 2012), but this solution affects the rest of the urban traffic. A way of solving the routing problem using genetic algorithms is by combining (Chand et al 2010) the Bin Packing (BPP) with the traveling salesman problem. In the most recent approaches, the solution to this problem was obtained using the Intelligent Water-drop algorithm (Kaur et al 2014).

There are multiple approaches to the application of multi-agent systems in dynamic reconfiguration of routes (Shah 2012), such as MARS, Jabatos or Ant System, but they were used for management problems of transport resources (e.g. assigning buses to routes by minimizing the number of vehicles required and maximizing the number of requests in the system).

One solution (Darbucha 2013) for the dynamic routing problem is a combination of agent-based systems and dispatcher-based routing strategies. Other similar approaches (Talbot et al 2010) using multi-criteria decision-making, but at a global level dispatcher. In this case, real time data is received by the dispatcher which computes the emergency vehicle route. If an obstacle is blocking this route, the dispatcher is informed and can take other decisions. This simulation is limited at dispatcher level by introducing an additional node between data processing and decision-making algorithm.

This paper is organized as follows. In section 2, the authors present the design of the proposed control system, along with the principles of genetic algorithms. In section 3 a case study is discussed, for which an agent based simulation model has been developed in order to simulate traffic in an urban area. Finally, section 4 contains the conclusions.

2 Proposed system design

The concept introduced in this paper focuses on obtaining routes for emergency crews in an urban environment. Urban conglomerates (especially in areas that had not been initially developed for the amount of vehicles that can be found in today's society) suffer from traffic congestions. Therefore, the shortest path between two points is not necessarily the fastest. In this paper, the objective is to find the shortest routes with the least amount of traffic at any moment, while also considering that the occupancy of a road segment might change during travel.

In order to achieve this goal, we propose the use of a Genetic Algorithm (GA) for computing the shortest of the least occupied routes through a network of intersections. The algorithm takes into account the degree of occupation for each intersection and chooses the route with the least amount of traffic.

Genetic Algorithms (GAs) are optimization heuristics able to perform rapid searches in large amounts of uncertain or incomplete data, with an inherent structure that allows parallelization. Given how GAs most often offer sub-optimal solutions is not a deterrent for the considered routing problem, because the urban traffic is in constant evolution and change. Thus, it is unadvisable to spend too much time and/or computing power on searching for a route that might not be as viable in the next minutes.

GAs are designed to search in multiple directions at once through the solution space, starting even from the initialization phase, in which a random pool of solutions (called a population) is generated. These solutions, known as individuals, are then allowed to evolve based on the principles of the Darwinist natural selection. Using mechanisms of selection (choosing certain individuals to contribute to creating the next generation as parents), recombination (generating new individuals by combining the characteristics of the selected parent), and mutation (necessary to maintain population diversity), a GA runs continuously, evaluating at each step all the possible solutions of the considered problem until a stop criterion is met (for instance, after a certain number of generations, or after reaching a certain tolerance for the solution). The evaluation procedure is run through what is known as a fitness function, that returns to which degree an individual can be deemed fit or not to be a viable solution.

Due to their heuristic nature, GAs can integrate, within the fitness function, objectives and performance indexes that are not necessarily defined in a formal manner. Moreover, adding more and more conditions in order to declare a possible individual as a fit solution to the considered problem is not computationally taxing (for instance, due to objective fusion techniques that might be hard to do with non-heuristic descriptions).

In recent research, GAs have been applied to problems in various fields. Their popularity keeps increasing due to their ease of use, effectiveness, or applicability. In control systems design (Fleming et al 2002) GAs have been efficiently used with both feedback (Lewin 1994) and feedforward controllers (Lewin 1996), modelling (Huang 2012) and parameter estimation (Bush et al 2011). In other fields, GAs have been used for various problems, either as stand alone procedures (Patrascu et al 2016), or in combinations

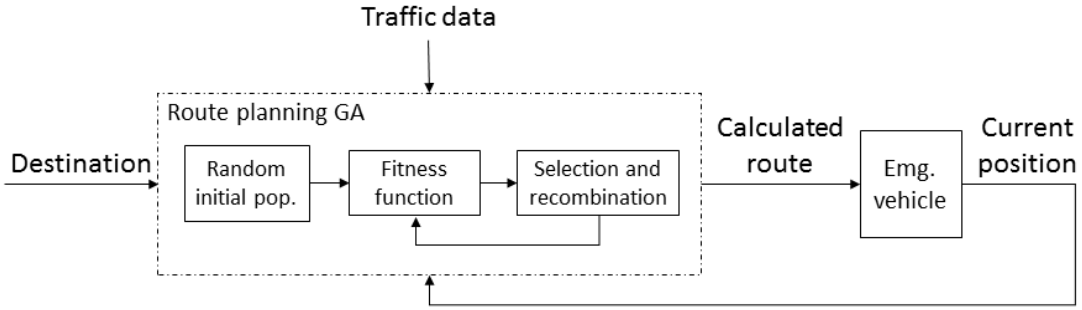


Figure 1. General principle of the dynamic route generation with genetic algorithms.

with other intelligent methods, like fuzzy systems (Alcala-Fdez et al 2009) or neural networks (Tao et al 2004).

In a real world implementation, the dynamic route generation algorithm would be developed to run independently onboard emergency vehicles, integrating traffic data from sensors throughout the city. It uses mobile computational resources available locally to the vehicle, thus eliminating the need for a central aggregation hub and offering scalability in case of major disasters, when more emergency crews are required in different parts of the city.

As the vehicle travels to the destination, the degree of occupation of the intersections along the route can change. In this case, the route is updated taking into account the new traffic data. The least occupied route is chosen based on the value of the corresponding fitness function (that computes a total degree of occupation for the entire suggested route).

Figure 1 presents the general principle of the proposed system. After a random initialization of the solution population, the GA computes the fitness function for each individual and, if necessary, genetic operators are applied to the population. The processes of evaluation, selection, and recombination are repeated until an individual with a high enough fitness is found or up to a maximum number of generations. Once a new route is generated by the GA, it is then transmitted to the driver.

At this point, whether the driver chooses to follow the suggested route or not, the GA starts its next run, in order to generate a new route by the time the vehicle reaches the next intersection along its path.

The entire system presented in figure 1 can be regarded as a control system in which the Genetic Algorithm (as *controller*) receives a destination (as a *setpoint*) and adjusts the trajectory (*controlled output*) of the emergency vehicle (*plant*) through a geographical area. The other traffic participants and the driver decision to ignore the suggested route are integrated into the control loop as *disturbances*. Each new suggested route is being transmitted to the driver at each node of the road network (intersection) as a *command*.

Thus, the control problem of emergency crew routes can defined as: an emergency vehicle (ambulance, fire truck, police car, etc.) receives a respond request to a given site while encountering as little traffic as possible during its journey. The emergency vehicle must travel from point A to point B, through an urban area comprised of a road network and traffic participants (regular vehicles and pedestrians).

The objective of the GA is to dynamically generate new routes for the emergency vehicle, supply these routes to the driver as the vehicle approaches each intersection, while taking into account traffic data that might be incomplete or inaccurate.

In what concerns the GA encoding of the vehicle routes, these are formed of a numeric representation in vectorial form, in which each position contains a route section identifier and the degree of occupation on the road segment leading to its associated intersection. The entire route is given by the position of each route section in the vector (its index).

Thus, an individual (or chromosome) is formed of genes that represent a route section each.

A route section is formed of that segment of a road between two intersections, on the traveling direction of the emergency vehicle. One road segment is associated with the intersection it travels towards, while the position of the emergency vehicle through a route is given by the road section & intersection pair it currently travels through. Thus, a path is formed of a starting point (either the initial position point A, or the current position), and as many route sections (road segment & intersection) as there can be delimited until the destination point B. For example, a path through 3 intersections will contain 4 genes.

A route section identifier can be defined as a coordinate pair or an unique identifier. This coordinate point can be, for instance, a physical geographical location for the beginning of each route section, a pair of coordinates for the start and finish of the route section, or even an ID number associated with each route section. For simplicity, in what follows, the

notation r will be used to denominate route section identifiers.

Thus, a route φ is comprised of n route sections defined by their position r_i in the urban area and their degree of occupancy D_i. Each pair $(r_i, D_i)_i$ has an associated index i that depicts their order in composing the route:

$$\varphi = [(r_0, D_0)_0 \quad (r_1, D_1)_1 \quad \dots \quad (r_i, D_i)_i \quad \dots \quad (r_n, D_n)_n] \quad (1)$$

When computing the fitness of an individual, the algorithm takes into account both the degree of occupation of the road segments on the possible paths to the event site and the distance the destination.

For a path formed of n intersections, the route fitness F is computed by analyzing each section and cumulating the results for the entire route:

$$F^{-1} = \sum_{i=1}^{n} \frac{\delta \cdot D_i}{g(r_0, r_i)} \quad (2)$$

where δ is a scaling factor, D_i is degree of occupancy on the i-th section of the route, and $g(r_0,r_i)$ is a function that scales the penalty given to those sections of the route that are farthest from the starting section r_0 (either the initial section or the current section the emergency vehicle is traveling through).

The fitness value F of the route needs to be maximized, therefore the best routes are the ones with the lowest degree of occupancy.

The closer a road section (delimited by the intersections through which the vehicle must travel), the more it is contributing to the computed fitness value than the further ones. In this way, when calculating the fitness of a possible route, less importance is given to the degree of occupancy of farther intersections, as it may change by the time the vehicle is actually there.

3 Case Study

In order to test the proposed system, a simulation model needs to be designed that incorporated the complexity of urban traffic systems. In this respect, the most recent development in complex systems modelling belong to the field of Agent Based Modelling and Simulation (Patrascu et al. 2015), which provides engineers with Agent Based Simulation Models (ABSM) that are able to describe the behaviours of heterogeneous traffic participants, as well as their interactions and interdependencies.

In an ABSM, an agent is a persistent entity characterised by internal states. This entity interacts with its environment or with other agents, ultimately causing changes in the environment or in the states of the other agents. An ABSM is a collection of agents, their states, and the rules that govern interactions from agents to agents, from agents to the environment, and from the environment to the agents.

Every agent of an ABSM has inputs and outputs, but it can also regulate their own internal states through local feedback loops. The internal state of an agent is governed by rules. Moreover, the behaviour of an agent is represented by the set of actions it performs (outputs) according to internal state and inputs.

As a whole, the ABSM has its own states, inputs, and outputs, but from a holistic perspective, the states of the entire simulation model integrates the complex interactions between the agents involved, by using rules to describe these behaviours. Most often, ABSM are evolving complex systems, used in generative experiments (Bertolotti, 2014), in order to test the viability of a solution that has little chance to be formally evaluated in realistic computing times (much like in the case of heuristics).

3.1 ABSM description

The agent-based simulation presented in this paper was implemented using the NetLogo (Wilensky et al 2015) environment, which offers access to the advantages of using agent based simulation models (ABSM) previously stated. Moreover, NetLogo offers way to implement user interfaces that is easy and accessible to researchers who don't necessarily have a strong programming background. NetLogo has been sucessfully used for proving and testing complex systems theorems and hypotheses. An in depth discussion on the advantages and disadvantages of using NetLogo for the simulation of complex control systems can be found in our previous work (Patrascu et al. 2015).

Thus, we illustrate the application of genetic algorithms in computing the quickest route (from the degree of occupation perspective) of an emergency vehicle through a simulated urban environment. The simulation application allows the user to set the number and types of crossroads present in the world model. Before starting the simulation, one can also chose the duration of different traffic light phases, the priorities for incoming traffic in a crossroad, as well as the rate at which new vehicles are inserted in the simulation. In this setup stage, the user can also chose the parameters that control the genetic algorithm responsible for planning the quickest route for the emergency vehicle.

For simulation purposes, the generated environment contains roads with four lanes (two lanes per travel direction) and specific traffic lights for turning left or going straight in order to increase the complexity of the road network model, as would be expected to happen in real urban areas: multiple lanes, directional traffic lights and so on. After starting the simulation, an emergency vehicle can be added to the environment. The parameters indicating the start and destination

position of the emergency vehicle, as well as its priority, can be set from the model's interface.

The traffic participants are simulated using mobile agents and consist of normal and emergency vehicles and pedestrians. The simulation environment, NetLogo, uses two types of agents: fixed and mobile. The fixed agents build the static parts of the simulation model, which, for the agent-based model compose their relevant environment, while the mobile agents are used to simulate dynamic entities. In this case, vehicles are mobile agents. Their behaviour is modelled using knowledge-based representations in a manner that most accurately describes the decision making processes of drivers. NetLogo offers the possibility of modelling these entities as complex or simplistic as the user desires. In this paper, the vehicle (and thus driver) behaviour is as close as possible to the behaviour of vehicles in real world traffic.

Emergency vehicles have a higher priority through traffic than civilian ones. The pedestrians behave similarly to the vehicles, with the exception that they can only move on the sidewalks. In figure 2 an example of simulated environment is presented, highlighting the intersection that the emergency vehicle (police) is currently traversing. The police car can be seen entering the inner crossroad area from the east and traveling west.

After creating the emergency vehicle, the user can chose to highlight and inspect the police car as it travels through the urban environment. This is implemented by opening a new window (presented in figure 2) that shows a zoom of the world model, centered on the emergency vehicle, and that follows it as it travels. In the zoom window presented here, we can see the state of the traffic lights in each intersection. The pedestrians obey the traffic lights placed in the corners of the intersection, while the vehicles follow the signals of the lights placed between the lanes. For vehicles, the traffic light closer to the center of the crossroad is for turning left and the other one for going straight. In figure 2, on the east-to-west road, we can see the traffic light for going straight showing a green light and the one for turning left showing red, thus allowing the police car to traverse the intersection and stopping the two civilian cars occupying the next lane from turning left. The route for the emergency vehicle is computed when the agent that models this vehicle is inserted in the simulation environment. The user can chose to have the route updated each time the vehicle crosses an intersection, or just request one calculation, at the beginning of the vehicle's journey.

During the simulation, at each time step, the phases of the traffic lights are computed for each crossroad, using a distributed traffic control system based on vehicle priority. Each traffic light phase (North, East, West, South) is computed using the corresponding road occupancy data, which is obtained from sensors placed on each road section entering the intersection. When a civilian vehicle enters a road section, the occupancy is increased by 1 and when the vehicle leaves the road section, it is decreased by the same amount. In the case of an emergency vehicle, the occupancy is modified according to the priority of the vehicle. For more information regarding these types of control systems and their simulation using ABSM, please consult our previous work (Patrascu et al 2015).

Figure 2. ABSM world overview.

As the simulation runs, at each time step, new vehicles are inserted in the simulation environment. The user can set the frequency for inserting vehicles from the east and west or north and south side of the map. The vehicles can only be inserted on a lane connected to the edge of the map if there is available room. Depending on a user setting, the civilian vehicles can give way to the emergency vehicles by temporary switching lanes.

3.2 A simulation example

For a more facile visualization, the world model presented in this case study has the road sections (segment & intersection pairs) displayed in a grid. In what follows, we have chosen a 4x4 grid of intersections. Each route section is depicted by an identifier of the form (x, y) where x is the position on each line of the grid, while y is the position on each column. For example, (1, 3) is the intersection marked A in figure 3. The starting point of the emergency vehicle (police car) is marked A in figure 3, while the final destination is B, at coordinates (4, 2).

For the first experiment (Exp.1), the GA presented in the previous sections will be run as the vehicle departs from point A. In the second experiment (Exp.2), the GA will be run at the entry of each route section, re-generating the vehicle's route dynamically

as it travels through the urban area. The results of both experiments are presented in table 1 (where D is a matrix containing the degrees of occupancy for each route section, and φ is the route returned by the GA). For Exp.1, the GA is run only at point A. For Exp.2, the table shows the first 2 runs of the GA, first in point A and then in the next intersection, showing how the occupancy of the road segments has changed while the vehicle has travelled through the first route section, from (1, 3) to (2, 3).

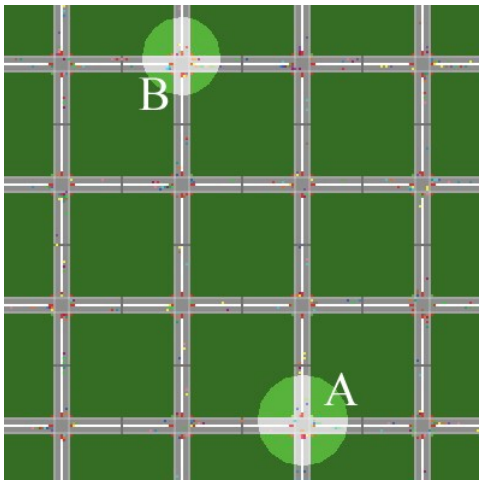


Figure 3. Starting point and destination of the emergency vehicle.

Table 1. Simulation run results

		Exp.1	Exp. 2
GA run at point A (1, 3)	*D*	$\begin{bmatrix} 34 & 8 & 7 & 37 \\ 40 & 10 & 12 & 46 \\ 38 & 14 & 10 & 52 \\ 18 & 16 & 9 & 36 \end{bmatrix}$	$\begin{bmatrix} 34 & 11 & 14 & 44 \\ 52 & 19 & 28 & 57 \\ 53 & 17 & 15 & 63 \\ 19 & 17 & 14 & 52 \end{bmatrix}$
	φ	 [(1,3), (2,3), (3,3), (4,3), (4,2)] Total Cars: 46	 [(1,3), (2,3), (2,2), (3,2), (4,2)] Total Cars: 76
GA run at point (2, 3)	*D*	$\begin{bmatrix} 44 & 18 & 14 & 58 \\ 51 & 20 & 29 & 42 \\ 50 & 18 & 28 & 61 \\ 29 & 31 & 10 & 44 \end{bmatrix}$	$\begin{bmatrix} 42 & 16 & 8 & 45 \\ 55 & 22 & 18 & 66 \\ 51 & 20 & 11 & 67 \\ 25 & 11 & 18 & 43 \end{bmatrix}$
	φ	 [(2,3), (3,3), (4,3), (4,2)] Total Cars: 89	 **New route:** [(2,3), (3,3), (4,3), (4,2)] Total Cars: 53  **Old route:** [(2,3), (2,2), (3,2), (4,2)] Total Cars: 69

In the first experiment, the vehicle's route is computed only once, at the beginning of its journey. As the vehicle travels through the area, the degree of occupancy might change on the previously selected routes. For instance, there is an alternate route starting at (2, 3) that has less vehicles traveling through it. To account for the changes in the environment, in the second experiment, the GA has been run at the beginning of each route section in order to determine if the route should be kept or dynamically changed. In table 1 a new and old route comparison is shown to illustrate this phenomenon.

The system we proposed in section 2 achieves its objectives of selecting the routes with the lowest degree of occupancy in a given urban area. Moreover, the system is capable of dynamically re-generating the emergency vehicle's route based on new traffic data and changes in the urban road network.

The preliminary testing conducted in this study shows great potential for the further development of evolutionary algorithms as controllers in closed loop systems. The next step in a thorough validation procedure is to develop simulation models that use real world maps, followed by real world testing. These are some of the main steps in the cycle of control systems development, which are designed to take a formal idea and make sure the final product works smoothly and within required performances.

4 Conclusions

In this study, a system for generating emergency vehicles routes has been presented. The system is able to account for changes in the environment or in urban traffic and dynamically supply new, better routes to the emergency vehicle drivers.

For this, a genetic algorithm has been designed that can recalculate the paths of the emergency responders while they are already on their journey to an event site. The entire route adjustment procedure is encapsulated in a control system scheme, in which the controlled variable is the route of the vehicle, while the other traffic participants are disturbances. The proposed system has been tested in a simulated environment specifically designed to emulate the complexity of the urban traffic.

Among the advantages of this type of system are: the possibility to run locally, on each emergency vehicle, thus eliminating the need for a central server type entity and overloaded communication networks; the possibility of implementation on devices that are already available (smartphones or tablets); not relying on traffic control systems, that can sometimes be overwhelmed, that can only manage an intersection at a time, or that can, at times, be offline (for maintenance or other reasons).

Moreover, for this type of system, context awareness is implicitly achieved. The intrinsic nature of control systems allows them to both collect information from the environment (either by direct sensing or by accessing real time traffic databases), and to interact with the it via the driver. Complete automation, although attractive from a point of view of vehicle autonomy, removes the driver from the decision making process. Thus, the middle ground we proposed seems the most reasonable.

Some of the limitations of this sort of system are related to the inherent heuristic nature of genetic algorithms, the proposed method requiring perhaps a sort of hybridization with formal methods, metaheuristics, memetic algorithms.

Further research endeavours in what concerns the proposed dynamic route generation using genetic algorithms include the design of specialized selection methods and chromosome encoding for routes and the test of the presented system, first in simulated environments with higher complexity, and then in real world scenarios.

Acknowledgements

We would like to thank our former student Eugen Marius Petre for his work within the EMAS (Emergent Multiscale Agents and Services) Research Group.

References

S. Andjelic, G. Panic, and A. Sijacki. Emergency response time after out-of-hospital cardiac arrest. *European journal of internal medicine*, 22(4): 386-393, 2011.

G. Asvin and G. Volker. Solving a dynamic real-life vehicle routing problem. *Operations research proceedings*: 367-372, 2006.

B. O. Bush, J.-P. Hosom, A. Kain, and A. Amano-Kusumoto. Using a genetic algorithm to estimate parameters of a coarticulation model. In *Twelfth Annual Conference of the International Speech Communication Association*, pages 2677-2680, 2011.

C.-F. Huang. A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing*, 12(2): 807-818, 2012.

D. R. Lewin. A genetic algorithm for MIMO feedback control system design. *Adv. Control Chem. Process*, 101: 2014, 1994.

D. R. Lewin. Multivariable feedforward control design using disturbance cost maps and a genetic algorithm. *Computers & chemical engineering*, 20(12): 1477-1489, 1996.

D. Barbucha. A multi-agent approach to the dynamic vehicle routing problem with time windows. In *International Conference on Computational Collective Intelligence,* pages 467-476, 2013.

G. B. Dantzig and J. H. Ramser. The truck dispatching problem. *Management science*, 6(1): 80-91, 1959.

G. F. Rushworth, C. Bloe, H. L. Diack, R. Reilly, C. Murray, D. Stewart, and S. J. Leslie. Pre-hospital ECG e-transmission for patients with suspected myocardial infarction in the highlands of Scotland. *International journal of environmental research and public health*, 11(2): 2346-2360, 2014.

H. N. Psaraftis. A dynamic programming solution to the single vehicle many-to-many immediate request dial-a-ride problem. *Transportation Science*, 14(2): 130-154, 1980.

H. N. Psaraftis. Dynamic vehicle routing: Status and prospects. *Annals of operations research*, 61(1): 143-164, 1995.

I. Benyahia and J.-Y. Potvin. Decision support for vehicle dispatching using genetic programming. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 28(3): 306-314, 1998.

J. White. Emergency vehicle priority. *The Queensland Surveying and Spatial Conference*, Brisbane Australia, 2012.

J. Barrachina, P. Garrido, M. Fogue, F. J. Martinez, J.-C. Cano, C. T. Calafate, and P. Manzoni. Reducing emergency services arrival time by using vehicular communications and Evolution Strategies. *Expert Systems with Applications*, 41(4): 1206-1217, 2014.

J.-Y. Potvin. A review of bio-inspired algorithms for vehicle routing. *Bio-inspired algorithms for the vehicle routing problem*, 1-34, 2009.

J. Alcalá-Fdez, R. Alcalá, M. J. Gacto, and F. Herrera. Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms. *Fuzzy Sets and Systems*, 160(7): 905-921, 2009.

M. M. Shah. Artificial Intelligence: Vehicle Routing Problem and Multi Agent System. *International Journal of Computer Applications*, DRISTI(1): 1-3, 2012.

M. Patrascu and A. Ion. Evolutionary Modeling of Industrial Plants and Design of PID Controllers. In H. E. Ponce Espinosa, editor, *Nature-Inspired Computing for Control Systems*, volume 40 of *Studies in Systems, Decision and Control*, pages 73-119, 2016.

M. Patrascu, A. Ion, and V. Constantinescu. Agent based simulation applied to the design of control systems for emergency vehicles access. In *ITS Telecommunications (ITST), 2015 14th International Conference on*, 50-54, 2015.

O. B. G. Madsen, H. F. Ravn, and J. M. Rygaard. A heuristic algorithm for a dial-a-ride problem with time windows, multiple capacities, and multiple objectives. *Annals of operations Research*, 60(1): 193-208, 1995.

P. Chand, B. S. P. Mishra, and S. Dehuri. A multi objective genetic algorithm for solving vehicle routing problem. *International Journal of Information Technology and Knowledge Management*, 2(2): 503-506, 2010.

P. Toth and D. Vigo. The vehicle routing problem, ser. SIAM monographs on discrete mathematics and applications. *Society for Industrial and Applied Mathematics*, 2002.

P. J. Fleming and R. C. Purshouse. Evolutionary algorithms in control systems engineering: a survey. *Control engineering practice*, 10(11): 1223-1241, 2002.

P. T. Pons, J. S. Haukoos, W. Bludworth, T. Cribley, K. A. Pons and V. J. Markovchick. Paramedic response time:

does it affect patient survival?. *Academic Emergency Medicine*, 12(7): 594-600, 2005.

Q. Tao, X. Liu, and M. Xue. A dynamic genetic algorithm based on continuous neural networks for a kind of non-convex optimization problems. *Applied mathematics and computation*, 150(3): 811-820, 2004.

R. Kaur, R. Kaur, and N. Kaur. A Modified transmission Algorithm for Resolving Vehicle Routing Problem by Intelligent Water drop Algorithm. *International Journal on Recent and Innovation Trends in Computing and Communication*, 2(10): 3108-3112, 2014.

R. Séguin, J.-Y. Potvin, M. Gendreau, T. G. Crainic, and P. Marcotte. Real-time decision problems: An operational research perspective. *Journal of the Operational Research Society*, 48(2): 162-174, 1997.

T. H. Blackwell and K. S. Jay. Response time effectiveness: comparison of response time and survival in an urban emergency medical services system. *Academic Emergency Medicine*, 9(4): 288-295, 2002.

T. Bertolotti. Generative and Demonstrative Experiments. In L. Magnani, editor, *Model-Based Reasoning in Science and Technology*, volume 8 of *Studies in Applied Philosophy, Epistemology and Rational Ethics,* pages 479-498, 2014.

U. Wilensky and W. Rand. *An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo*, MIT Press, 2015.

V. Talbot and I. Benyahia. Complex Application Architecture Dynamic Reconfiguration Based on Multi-criteria Decision Making. *International Journal of Software Engineering & Applications*, 1(4): 19-37, 2010.

W.-R. Jih and J. Y.-J. Hsu. Dynamic vehicle routing using hybrid genetic algorithms. In *Proceedings of the 1999 IEEE International Conference on Robotics and Automation*, *Detroit, USA*, pages 453-458, 1999.

The Effect of Pressure Losses on Measured Compressor Efficiency

Kristoffer Ekberg Lars Eriksson

Vehicular Systems, Linköping University, Sweden, {kristoffer.ekberg, lars.eriksson}@liu.se

Abstract

While measuring the compressor behavior at different load points in for example a gas stand, the inlet and outlet pressures are not always measured directly before and after the compressor. The friction inside the pipes and the physical piping configuration affect the measured compressor efficiency, due to the induced change of fluid enthalpy. If the measured pressures at the end of the inlet and outlet pipes are not the same as the actual pressure before and after the compressor, the acquired compressor map does not give the right description of it as an isolated component. The main contribution of this paper is the analysis of the impact of gas stand energy losses due to pipe friction on the compressor map. As a result the paper suggests a way to take the pressure losses in the inlet and outlet pipes into account. The suggested model takes pipe friction, diffuser, nozzle and pipe bends into account. The potential measurement error in compressor efficiency due to energy losses in the pipes in this experiment is 2.7% (percentage points) at maximum mass flow of air through the compressor.

Keywords: gas stand, pipe, bend, diffuser

1 Introduction

Gas stand testing of turbochargers is a time consuming process where one of the goals is to determine the compressor efficiency. Turbochargers are modeled in computers to perform more cost efficient tests and experiments. Softwares today are used to solve and compute the dynamic behaviors of complex engine systems involving turbochargers. Turbocharger models are often adjusted to fit measured data, from for example a gas stand test. Most of the analysis assume that the turbocharger models represent the turbocharger as a single component. This means that to have accurate turbocharger models, the measurement data should represent the turbocharger only, and not include any pipes or other objects connected to the turbocharger housing. The pipes in a gas stand, connecting the turbocharger to measurement instruments, induces errors into the computer models if the data is used without correction. Since the pressures and temperatures are measured some distance away from the actual inlet and outlet on the compressor, the physical setup of the gas stand may need to be accounted for to get a more accurate result of the compressor efficiency. In both the inlet and outlet pipes there are pressure losses due to friction inside the pipes, also if the gas flow path contains bends or area changes, these could induce pressure losses. There are different ways to develop a gas stand (see for example (Venson et al., 2006) or (Young and Penz, 1990)), the idea is to simulate engine conditions to find the turbocharger characteristics. When making measurements in a gas stand, the monitoring of the pressures and temperatures before and after the compressor are important to get accurate results of the compressor efficiency (Kumar et al., 2014). The compressor efficiency is determined by using measured values of temperatures and pressures before and after the compressor. Studies with focus on the heat transfer inside the turbocharger (Nick Baines and Karl D.Wygant and Antonis Dris, 2009) and how the heat transfer affects the compressor efficiency have been performed (Marelli et al., 2015), while others focusing on the possible measurement errors due to sensor inaccuracy (Guillou, 2013). No papers are found where the actual placement of the sensors are examined up or downstream from the compressor, SAE standard J1826 recommends placing the static pressure taps 2 to 3 pipe diameters downstream of the rotor (SAE, 1995). The sensor placement is crucial to achieve a reliable result during testing. Different test rigs may give different results due to environmental conditions, if the inlet air is not conditioned, the efficiency uncertainty will fluctuate (Guillou, 2013). The impact from inlet air being dry or humid on compressor efficiency has been studied in (Serrano et al., 2009), the impact is small and should only be considered if very high accuracy is wanted. The enthalpy loss between the measurement positions and the compressor due to the pressure loss indicates that the compressor efficiency is actually better than measured.

1.1 Contributions

This paper is the first to analyze the gas stand pressure losses. The effects of the pressure losses on measured compressor efficiency are analyzed and ways to compensate for them are developed. Influences of the gas stand pressure losses are displayed on the compressor map.

1.2 Setup for the analysis

The main scope of the paper is to show how the change of enthalpy in the inlet and outlet pipes affect the compressor efficiency. The inlet and outlet enthalpies ($\dot{h}_{in}$ and $\dot{h}_{out}$ in Figure 1) are not the enthalpies actually entering and leaving the turbocharger compressor, the actual values are $\dot{h}'_{in}$ and $\dot{h}'_{out}$, which are corrected to exclude the friction in the inlet and outlet pipes ($\dot{w}_{f,pipe,in}$ăand $\dot{w}_{f,pipe,out}$). In

equation (1a) the work used by the compression process is described as it is used today, where the connecting inlet and outlet pipes are included, since the measurement positions are not located directly at the inlet or outlet of the compressor. Equation (1b) describes the work required by the compressor as a single component, describing the compression work made by the turbocharger compressor only.

$$\dot{w}_{c,Old} = \dot{w}_{compression} + \dot{w}_{f,pipe} + \dot{w}_{f,pipe} \quad (1a)$$

$$\dot{w}_{c,New} = \dot{w}_{compression} \quad (1b)$$

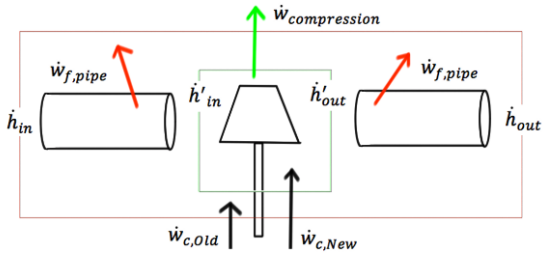


Figure 1. Energy flow in the compressor. The outer box (red) represents the system that is measured in a gas stand, the inner box (green) is the preferred system that is to be described by the model.

To quantify the impact from the inlet and outlet pipe frictions ($\dot{w}_{f,pipe,in}$ and $\dot{w}_{f,pipe,out}$) on the measured compressor efficiency, three cases are investigated. The three cases investigates:

1. straight inlet and outlet pipes, see Figure 2. The diameter of the pipes (d_{inlet} and d_{outlet}) are assumed to be equal to the compressor inlet and outlet diameters ($d_{c,inlet}$ and $d_{c,outlet}$).
2. using pipes with diffuser and nozzle. See Figure 3.
3. adding a 90^o smooth bend on the inlet pipe. See Figure 4.

1.3 Experimental data

Data used during the analysis is a measured compressor map from a commercial turbocharger, the measured mass flow range [0, 0.21] kg/s and pressure ratio between [1, 2.8] $\frac{p_{02}}{p_{01}}$. The sensor errors effect on the achieved results from a gas stand have been studied in (Guillou, 2013). The measured data used in this analysis are assumed to be correct, i.e. all measured values are assumed to be perfect, no sensor errors are assumed to be present.

1.4 Compressor Map

One of the main ideas behind testing the turbocharger in a gas stand is to determine the compressor efficiency and flow characteristics at different work points. The compressor behavior is presented on a compressor map, where the corrected compressor mass flow and pressure ratio defines

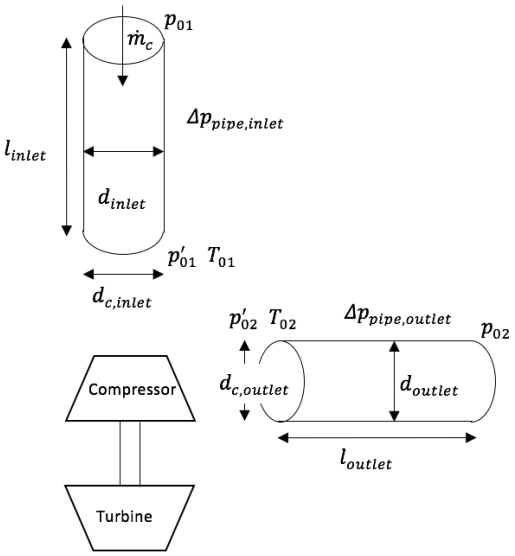


Figure 2. Pressure drops are represented by Δp_n, the total pressures by p_{01} and p_{02}, the corrected total pressures by p'_{01} and p'_{02} and the measured temperatures by T_{01} and T_{02}. The mass flow of air inside the pipes are represented by $\dot{m}_c$. The physical dimensions on pipe lengths and pipe diameters are described by l_n and d_n.

a plane where the compressor efficiency is displayed. In the evaluation of the results, the effects from the pressure losses on compressor efficiency are presented on the compressor map. The reference compressor efficiency is calculated using measured data, and later recalculated when taking the pressure losses in the gas stand into account.

1.5 Compressor Isentropic Efficiency

The compressor isentropic efficiency is defined as the smallest amount of power needed to compress the air without heat exchange with the environment (isentropic process), divided by the actual amount of power consumed by the process. Using measured total temperatures and total pressures (calculated from static pressures, see equation (4)) from a gas stand, the compressor total to total isentropic efficiency can be calculated using equation (2). (Eriksson and Nielsen, 2014)

$$\eta_c = \frac{\Pi_c^{\frac{\gamma-1}{\gamma}} - 1}{\frac{T_{02}}{T_{01}} - 1}, \text{ where } \Pi_c = \frac{p_{02}}{p_{01}} \quad (2)$$

where p_{01} and p_{02} are total pressures, T_{01} and T_{02} are total temperatures and γ is the ratio of specific heats (assumed to be constant).

1.6 Corrected Mass Flow

Corrected mass flow is used to display the mass flow in the compressor map. The corrected mass flow is used instead of the measured mass flow, to take surrounding conditions during measurements into account. The surrounding conditions are the reference temperature, T_{ref} and the refer-

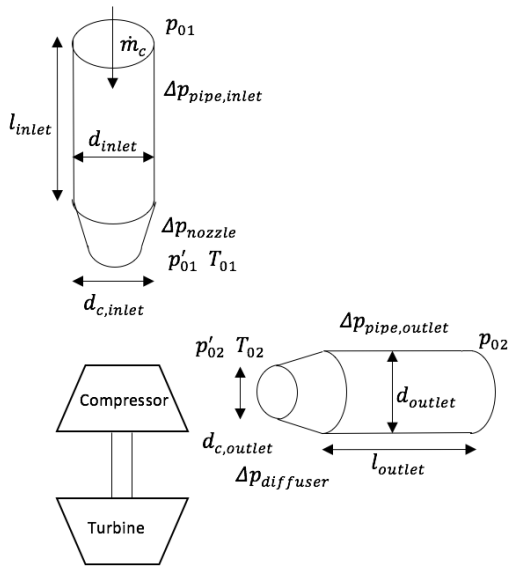


Figure 3. Pressure drops are represented by Δp_n, the total pressures by p_{01} and p_{02}, the corrected total pressures by p'_{01} and p'_{02} and the measured temperatures by T_{01} and T_{02}. The mass flow of air inside the pipes are represented by $\dot{m}_c$. The physical dimensions on pipe lengths and pipe diameters are described by l_n and d_n.

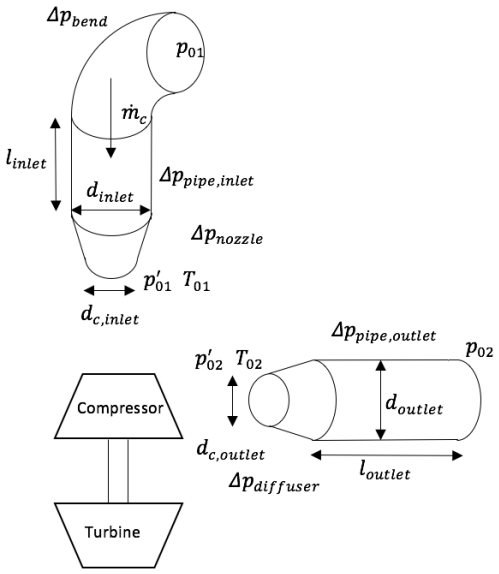


Figure 4. Pressure drops are represented by Δp_n, the total pressures by p_{01} and p_{02}, the corrected total pressures by p'_{01} and p'_{02} and the measured temperatures by T_{01} and T_{02}. The mass flow of air inside the pipes are represented by $\dot{m}_c$. The physical dimensions on pipe lengths and pipe diameters are described by l_n and d_n.

ence pressure, p_{ref}. The corrected mass flow is calculated according to equation (3). (Eriksson and Nielsen, 2014)

$$\dot{m}_{c,corr} = \frac{\dot{m}_c \sqrt{\frac{T_{01}}{T_{ref}}}}{\frac{p_{01}}{p_{ref}}} \tag{3}$$

1.7 Data Treatment

The pressures and temperatures used when calculating the compressor efficiency should be the total pressures and total temperatures. The relation between the measured static pressure p_i and the total pressure p_{0i} is displayed in equation (4), where $C_i = \frac{\dot{m}_c}{\rho_i A}$ is the fluid velocity inside pipe i.

$$p_{0i} = p_i + \frac{\rho_i C_i^2}{2} \tag{4}$$

To convert the measured total temperature T_{0i} to static temperature T_i equation (5) from (Eriksson and Nielsen, 2014) can be used.

$$T_i = \frac{A^2 p_i^2 c_p}{R^2 \dot{m}_c^2} \left(\sqrt{1 + 2\frac{R^2 \dot{m}_c^2 T_{0i}}{A^2 p_i^2 c_p}} - 1 \right) \tag{5}$$

where $\dot{m}_c$ is the mass flow of air through the pipe, ρ_i is the air density and A is the cross section area of the pipe at the measurement position. The air inside the system is treated as an ideal gas, following this assumption, the density is calculated using equation (6). (Eriksson and Nielsen, 2014)

$$\rho_i = \frac{p_i}{RT_i} \tag{6}$$

2 Pressure Losses in Gas Stand

Pressure losses in different piping systems and pipe configurations have been examined for many years. The formulas and expressions are empirical or semi-empirical correlations that are created from experiments to describe specific objects or system configurations. The different pressure losses in different parts in the gas stand are calculated according to empiric formulas, these formulas are valid for fully developed turbulent flow, the turbulent flow in the parts taken into account is therefor assumed to be fully developed. Both the total temperature and the density of the fluid are assumed to be constant along the inlet and outlet pipe sections.

2.1 Pressure Loss in Straight Pipe

Straight pipes in for example a gas stand causes pressure losses due to friction inside the pipes. The selection of pipe material and manufacturing method of the pipes are important to get a low friction pipe. The surface roughness inside the pipe induces pressure loss when the flow is turbulent, when the flow is laminar, the friction factor $f_{pipe,i}$ is independent of the surface roughness. The pressure loss in a straight pipe is calculated with equation (7) (Cengel et al., 2008).

$$\Delta p_{pipe} = f_{pipe,i} \frac{l_i}{d_i} \frac{\rho_i v_i^2}{2} \tag{7}$$

Where $f_{pipe,i}$ is a friction factor, ρ_i is the density of the fluid inside the specific pipe section i, l_i is the pipe section length, v_i is the mean velocity of the fluid inside the

specific pipe section i. The friction factor $f_{pipe,i}$ is dependent on the flow characteristics inside the pipe. The flow characteristics could be either laminar or turbulent. The flow characteristics inside the pipes are determined by Reynolds number, Re. Reynolds number is calculated according to equation (8). (Cengel et al., 2008)

$$Re = \frac{v_i d_i \rho_i}{\mu_i} \tag{8}$$

Where v_i is the mean velocity of the fluid inside the pipe, ρ_i is the fluid density, d_i is the hydraulic diameter (hydraulic diameter equals pipe diameter for circular pipes) and μ_i is the dynamic viscosity of the fluid. The dynamic viscosity of air is described as a function of air temperature:

$$\mu_i = \mu(T_i) \tag{9}$$

According to (White, 1999), the change in μ is around 10% for air when the pressure is increased from 1 to 50 atm, and that it is customary in most engineering work to neglect the pressure variations. The viscosity of a gas is by (Massey and Ward-Smith, 1998) said to be independent of its pressure (except at very high or very low pressures). In this study, the change in pressure ranges from around 1 bar to 2.85 bar, therefore the fluid dynamic viscosity is assumed to be independent of the pressure variations. The function in equation (9) describes the fluid dynamic viscosity $\mu(T_i)$, as a polynomial function of fluid temperature T_i, the function parameters are adapted to fit data from table A-22 in (Cengel et al., 2008) (Properties of air at 1 atm pressure), the function is displayed in equation (10).

$$\mu(T_i) = -3.0777 \times 10^{-11} T_i^2 + 4.8218 \times 10^{-8} T_i + 1.7299 \times 10^{-5} \tag{10}$$

For low Re, the flow is considered to be laminar, for higher Re, the flow is considered to be turbulent. In-between the laminar and turbulent region there is a region where the flow is called transitional flow. When the flow is transitional, the flow is frequently shifting between laminar and turbulent. The limits on Re is shown in equation (11). (Cengel et al., 2008)

$$\begin{cases} Re \leq 2300 & \text{Laminar flow} \\ 2300 < Re < 10000 & \text{Transitional flow} \\ Re \geq 10000 & \text{Turbulent flow} \end{cases} \tag{11}$$

Laminar and turbulent flow are the two flow characteristics that will be taken into account. The flow is mostly turbulent during the gas stand test performed, but the laminar region will be described to make the model complete.

2.2 Friction Factor - Laminar Flow

To calculate $f_{pipe,i}$ when the flow is laminar, equation (12) is used. (Cengel et al., 2008)

$$f_{pipe,i} = \frac{64}{Re} \tag{12}$$

2.3 Friction Factor - Turbulent Flow

During turbulent flow inside the pipe, the surface roughness of the pipe ε affects the pressure loss inside the pipe (assuming pipe material to be stainless steel with surface roughness $\varepsilon = 0.002mm$ from table 14-1 in (Cengel et al., 2008) during calculations). To calculate $f_{pipe,i}$ when the flow is turbulent, either equation (13), known as Colebrook equation, is used and iterated until $f_{pipe,i}$ is accurate enough, or equation (14) could be used. The result of equation (14) is within 2 % of the result from equation (13). (Cengel et al., 2008)

$$\frac{1}{\sqrt{f_{pipe,i}}} = -2.0 log\left(\frac{\varepsilon/d_i}{3.7} + \frac{2.51}{Re\sqrt{f_{pipe,i}}}\right) \tag{13}$$

$$\frac{1}{\sqrt{f_{pipe,i}}} \cong -1.8 log\left(\frac{6.9}{Re} + \left(\frac{\varepsilon/d_i}{3.7}\right)^{1.11}\right) \tag{14}$$

2.4 Pressure Loss In Bend

Pipe bends are treated as one-time losses, a smooth 90^o bend has a loss coefficient of $K_L = 0.3$. The value of K_L is strongly dependable on the type of pipe, size of bend etc., the coefficient value is found in table 14-3 in (Cengel et al., 2008), it is used to give a hint about how the losses affect the measured compressor efficiency. The pressure drop due to a pipe bend is calculated according to equation (15).

$$\Delta p_{bend} = \frac{K_L v_i^2 \rho_i}{2} \tag{15}$$

2.5 Pressure Loss in Inlet Nozzle and Outlet Diffuser

Inlet nozzle and outlet diffuser can be used to connect the inlet and outlet pipes to the turbocharger. The inlet nozzle is treated as a convergent pipe, a convergent pipe is not inducing any pressure loss over the area change, other than the friction in the pipe. This is due to the contraction of the pipe, a gradually contracting pipe is normally not inducing any extra turbulence, other pressure losses than the pipe friction is normally neglected (Nakayama and Boucher, 1999), the pressure loss in the nozzle is neglected in this study (see equation (16b)). The friction loss in the inlet nozzle is assumed to be included in the pressure loss inside the inlet pipe. The outlet diffuser induces a pressure loss, due to the extra turbulence induced in the divergent region. The pressure drop in the outlet diffuser is calculated according to equation (16a), the pressure drop due to pipe friction is assumed to be included in the expression.

$$\Delta p_{diffuser} = \frac{K_{L,exp} v_i^2 \rho_i}{2} \tag{16a}$$

$$\Delta p_{nozzle} = 0 \tag{16b}$$

The value of $K_{L,exp}$ is found using tables, the used value is found in table 14-3 in (Cengel et al., 2008) (assumption of the diffuser angle 20^o results in $K_{L,exp} = 0.1$, when $d_{c,outlet}/d_{outlet} = 0.8$). The fluid velocity inside the pipe v_i is the fluid velocity at the diffuser inlet.

2.6 Adjust Measured Data

The measured data is adjusted by summarizing and withdrawing the pressure losses in the gas stand, from the measured values. The adjustments made are shown in equation (17a) and (17b). The inlet and outlet total pressures are adjusted by adding or subtracting the pressure losses, depending on if the losses occur up or downstream from the measurement positions.

$$p'_{01} = p_{01} - \Delta p_{pipe,inlet} - \Delta p_{bend} - \Delta p_{nozzle} \quad (17a)$$

$$p'_{02} = p_{02} + \Delta p_{pipe,outlet} + \Delta p_{diffuser} \quad (17b)$$

2.7 Calculate New Compressor Efficiency

The new corrected compressor efficiency is calculated using the total pressures that are adjusted to measurement data (see equation (17a) and (17b)) and the measured total temperatures. The equation to calculate the corrected efficiency is the same as equation (2), but with the new corrected pressures p'_{01}ăand p'_{02} (see equation (18)).

$$\eta'_c = \frac{\left(\frac{p'_{02}}{p'_{01}}\right)^{\frac{\gamma-1}{\gamma}} - 1}{\frac{T_{02}}{T_{01}} - 1} \quad (18)$$

3 Effect of pressure losses on measured compressor efficiency

Different simulation cases are performed to quantify the pressure losses main impacts on the measured compressor efficiency. The first case investigates the usage of straight inlet and outlet pipes, with the same diameter as the compressor inlet and outlet. The second case studies the usage of nozzle and diffuser on the inlet and outlet pipe, to connect a larger inlet and outlet pipe to the compressor. The third case is the same as the second, but a 90^o bend is added on the inlet pipe.

3.1 Compressor and Pipes Dimensions

A measured compressor map from a gas stand test is used to quantify the error in compressor efficiency due to the pressure drop between measurement positions and the turbocharger compressor. Compressor inlet outlet diameters and diameters at measurement locations are displayed in Table 1, these dimensions are needed to calculate the pressure drops in the different pipe sections. The results in Table 2 shows the maximum pressure loss over the different components. The inlet pipe is assumed to be 100 mm long and the outlet pipe is 3 or 10 times the outlet pipe diameter, both inlet and outlet pipe diameters are assumed to be the same as compressor inlet and outlet diameter when analyzing Case 1. According to SAE standard J1826 (SAE, 1995), the distance from the rotor down to the measurement location (if measuring static pressure) should be 2 to 3 pipe diameters downstream. In many pipe flows of practical engineering interest, the effects due to the entrance region become insignificant when the pipe length is longer than 10 pipe diameters (Cengel et al., 2008). Two different selections of outlet pipes lengths (10 and 3 times the outlet pipe diameter) are studied and compared in terms of measurement error due to the simulated pressure losses.

Table 1. Diameter of compressor inlet and outlet on the turbocharger, diameter on the inlet and outlet pipe at the measurement positions. Pipe lengths are assumed.

Measurement	Value
$d_{c,inlet}$ (inlet compressor)	56.5 *mm*
d_{inlet} (measurement position p_{01})	58 *mm*
$d_{c,outlet}$ (outlet compressor)	40 *mm*
d_{outlet} (measurement position p_{02})	50 *mm*
l_{inlet}	100 *mm*
l_{outlet}	
3 $d_{c,outlet}$ or 10 $d_{c,outlet}$	(case 1)
3 d_{outlet} or 10 d_{outlet}	(case 2, 3)

Table 2. Maximum pressure loss in the different components, for all 3 cases. The pressure losses are presented in Pascal.

Case	Δp_{inlet}	Δp_{outlet}	$\Delta p_{diff.}$	Δp_{bend}	Δp_{nozzle}
1	72	305	0	0	0
1	72	1016	0	0	0
2	71	315	1637	0	0
2	71	1050	1637	0	0
3	71	315	1637	803	0
3	71	1050	1637	803	0

3.2 Case 1: Straight Pipes

The first case investigates the use of straight inlet outlet pipes. Figure 5, Case 1, show the change in compressor efficiency for both short and long outlet pipe. In figure, it is visual that the pressure losses does not affect the compressor efficiency more than 0.8% (percentage points) at maximum mass flow of air when analyzing the long outlet pipe.

3.3 Case 2: Pipes with Diffuser and Nozzle

The diffuser and nozzle are used to either increase or decrease fluid pressure or velocity. The simulated inlet and outlet pipes with the pressure sensors mounted are assumed to have the same diameter as the pipe at the static pressure sensor location. The size of the diffuser is chosen to connect the pipe diameter where the pressure measurement is made, and the diameter of the compressor outlet. The results for both short and long outlet pipes are displayed in Figure 5, Case 2. Since the pressure loss

caused by the nozzle is assumed to be zero, the pressure loss caused by the diffuser and the pipe friction causes the efficiency calculation error to be 1.4% (percentage points) when analyzing the short outlet pipe and up to 1.9% when analyzing the long outlet pipe.

3.4 Case 3: Pipes with Diffuser, Nozzle and Bend

A 90^o smooth bend is added to the simulated inlet pipe, between the pressure sensor and the pipe connecting to the compressor, to find its impact on the measured compressor efficiency, see Figure 5, Case 3. The pipe bend clearly affects the results, this is visual if comparing Case 2 with Case 3. The maximum error in the calculated compressor efficiency is 2.7% (percentage points), when analyzing the long outlet pipe.

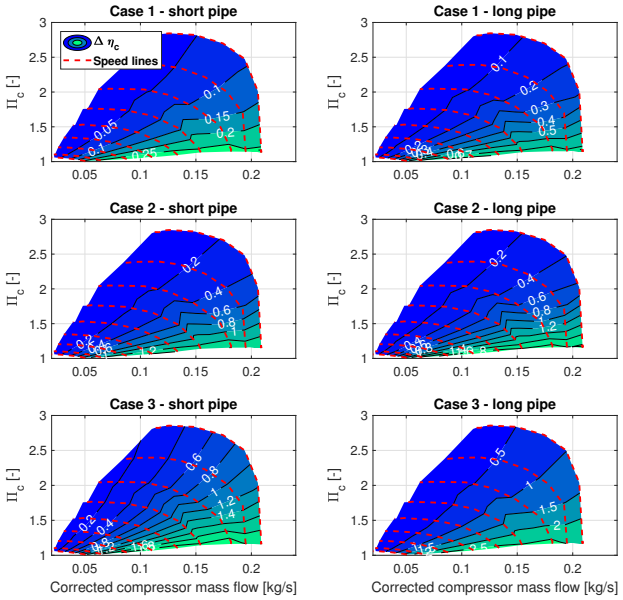


Figure 5. Displays the change in compressor efficiency $\Delta\eta_c = \eta_c' - \eta_c$ (color scale) compared against Π_c and $\dot{m}_{c,corr}$. Short pipes corresponds to outlet pipe length equal to 3 times the pipe diameter, long pipes corresponds to 10 times the pipe diameter.

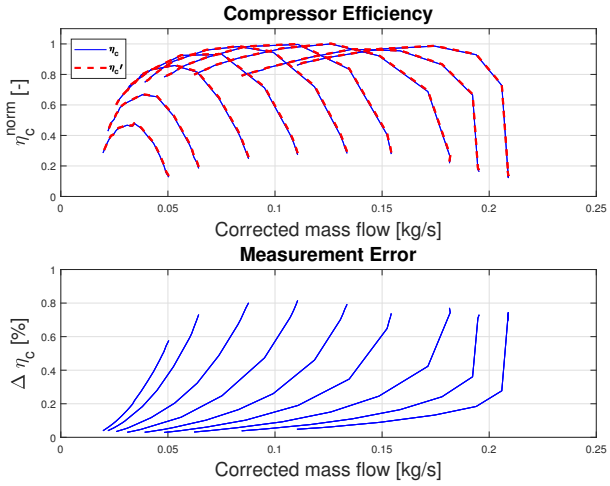


Figure 6. Case 1, the pipe length is 10 times the outlet pipe diameter. Top figure shows the normalized compressor efficiency with and without correction for the pressure losses in the gas stand, the bottom figure shows $\Delta\eta_c = \eta_c' - \eta_c$.

4 Summary and Discussion

Three different cases have been investigated to find and quantify the error in compressor efficiency due to enthalpy change in the inlet and outlet pipes. The enthalpy change present between the pressure sensors and the compressor affects the compressor map the most in the high flow low pressure region, for each speed line. This is visible in both Figure 5 and Figure 6. For all the displayed cases, the error in compressor efficiency increases with increasing mass flow. Case 1 shows that a longer pipe between the compressor outlet and the measurement location induces larger error in measured compressor efficiency. Case 2 studies the use of pipes with nozzle and diffuser, the nozzle is assumed to not induce any pressure loss, which shows that the diffuser induces a large pressure loss, which affects the compressor efficiency. Comparing Case 2 with Case 3, where the difference is the introduced pipe bend, clearly shows that a pipe bend induces errors in the calculated compressor efficiency. If a pipe bend is present between the pressure sensor and the compressor, it should be taken into account to correct measurements. The magnitude of the pressure losses in Table 2 are small, but they still affect the compressor efficiency noticeably.

5 Future Work

If this study is to be extended, one interesting aspect would be to investigate the impact on engine performance, if the compressor efficiency is corrected according to the study. The study could also be extended to take the heat transfer inside the inlet and outlet pipes into account.

6 Conclusions

For the selected set of gas stand physical dimensions, the change in compressor efficiency due to the calculated

pressure losses is compared to original compressor efficiency calcuations. The results show:

- Due to the friction work, the enthalpy of the fluid between the pressure sensors and the compressor inlet and outlet changes.
- The measured compressor efficiency is lower than the actual efficiency, due to pressure losses between compressor and the pressure sensors.
- The induced error $\Delta\eta_c$ shows that the error is getting larger with increased mass flow for each speed line.
- If a 90^o bend is present between the measurement position and the inlet to the compressor, and the diffuser is connected on the outlet, the error in calculated compressor efficiency is up to 2.7% for compressor maximum mass flow with used parameters.
- The pressure losses in the inlet and outlet pipes are affecting the compressor efficiency most at the high flow low pressure region for each speed line, where the compressor efficiency is generally low.

Acknowledgment

This work was supported by the Vinnova Industry Excellence Center: LINK-SIC Linköping Center for Sensor Informatics and Control.

References

SAE Standard. J1826 - Turbocharger Gas Stand Test Code, 1995.

Yunus A. Cengel, Robert H. Turner, and John M. Cimbala. *Fundamentals of Thermal-Fluid Sciences*. McGraw-Hill, Singapore, 2008.

Lars Eriksson and Lars Nielsen. *Modeling and Control of Engines and Drivelines*. John Wiley and Sons Ltd, United Kingdom, 2014.

Erwann Guillou. Uncertainty and measurement sensitivity of turbocharger compressor gas stands. In *SAE Technical Paper*. SAE International, 04 2013. doi:10.4271/2013-01-0925. URL http://dx.doi.org/10.4271/2013-01-0925.

Sathvick Shiva Kumar, Bert van Leeuwen, and Aaron Costall. Quantification and Sensitivity Analysis of Uncertainties in Turbocharger Compressor Gas Stand Measurements Using Monte Carlo Simulation. In *SAE Technical Paper*. SAE International, 04 2014. doi:10.4271/2014-01-1651. URL http://dx.doi.org/10.4271/2014-01-1651.

Silvia Marelli, Giulio Marmorato, Massimo Capobianco, and Andrea Rinaldi. Heat transfer effects on performance map of a turbocharger compressor for automotive application. In *SAE Technical Paper*. SAE International, 04 2015. doi:10.4271/2015-01-1287. URL http://dx.doi.org/10.4271/2015-01-1287.

Bernard Massey and John Ward-Smith. *Mechanics of Fluids*. Stanley Thornes, United Kingdom, 1998.

Y. Nakayama and R.F. Boucher. *Introduction to Fluid Mechanics*. Arnold, Great Brittan, 1999.

Nick Baines and Karl D.Wygant and Antonis Dris. The Analysis of Heat Transfer in Automotive Turbochargers. International Gas Turbine Institute of ASME, 2009. doi:10.1115/1.3204586.

J. R. Serrano, V. Dolz, A. Tiseira, and A. Páez. Influence of environmental conditions and thermodynamic considerations in the calculation of turbochargers efficiency. In *SAE Technical Paper*. SAE International, 04 2009. doi:10.4271/2009-01-1468. URL http://dx.doi.org/10.4271/2009-01-1468.

Giuliano Gardolinski Venson, Jose Eduardo Mautone Barros, and Josemar Figueiredo Pereira. Development of an automotive turbocharger test stand using hot gas. In *SAE Technical Paper*. SAE International, 11 2006. doi:10.4271/2006-01-2680. URL http://dx.doi.org/10.4271/2006-01-2680.

Frank M. White. *Fluid Mechanics*. McGraw-Hill, Singapore, 1999.

Michael Y. Young and David A. Penz. The design of a new turbocharger test facility. In *SAE Technical Paper*. SAE International, 02 1990. doi:10.4271/900176. URL http://dx.doi.org/10.4271/900176.

Implementation of an Optimization and Simulation-Based Approach for Detecting and Resolving Conflicts at Airports

Paolo Scala [1] Miguel Mujica Mota [2] Daniel Delahaye [3]

[1,2]Aviation Academy, Amsterdam University of Applied Sciences, The Netherlands, {p.m.scala,m.m.mujica}@hva.nl

[3]Ecole Nationale de l'Aviation Civile, delahaye@recherche.enac.fr

Abstract

In this paper is presented a methodology that uses simulation together with optimization techniques for a conflict detection and resolution at airports. This approach provides more robust solutions to operative problems, since, optimization allows to come up with optimal or suboptimal solutions, on the other hand, simulation allows to take into account other aspects as stochasticity and interactions inside the system. Both the airport airspace (terminal manoeuvring area), and airside (runway taxiways and terminals), were modelled. In this framework, different restrictions such as speed, separation minima between aircraft, and capacity of airside components were taken into account. The airspace was modeled as a network of links and nodes representing the different routes, while the airside was modeled in a low detail, where runway, taxiways and terminals were modeled as servers with a specific capacity. The objective of this work is to detect and resolve conflicts both in the airspace and in the airside and have a balanced traffic load on the ground.

Keywords: optimization, modeling, simulation, airport

1 Introduction

Capacity at airports has become a very delicate problem due to the increase of traffic demand and the scarcity of facilities at airports. In Europe it has been seen a growth of traffic of 1.5\% from 2014 to 2015, and the forecasts say that this growth will continue also for the coming years (Eurocontrol, 2016). Airports are getting busier and busier, especially at the major hubs in Europe, with visible effects as delays occurrences. Looking at the delay from all causes, it can be seen that in the first three months of the year there were between 34\% and 38\% flights delayed on departures, where only delays greater or equal than five minutes are considered (Eurocontrol, 2016). So far, many studies have been conducted in order to alleviate airports from congestion and improve the capacity, some of them focused on the airspace and some other only on the airside. Concerning problems related to the airspace, we can find many works about the sequencing and merging or scheduling of arrivals in the TMA (Beasley et al., 2000; Beasley et al., 2001; Hu and Chen, 2005; Michelin et al., 2009; Balakrishnan and Chandran, 2010; Zuniga et al. 2011). On the other hand we can also find studies related to ground issues such as gate assignment problem (Bolat, 2000; Dorndorf et al., 2007; Kim and Feron, 2012; Narciso and Piera, 2015), scheduling of departures (Pujet et al., 1999; Rathinam et al. 2009; Sandberg et al., 2014; Simaiakis and Balakrishnan, 2015) or airport surface management (Montoya et al., 2010; Simaiakis et al., 2014; Khadilkar and Balakrishnan, 2014). In order to conduct a more precise analysis and obtaining an integrated view of the system, it is better to consider airspace and airside together, so that we can also consider the interactions between the two environments.

The contribution of this paper is that it considers the airport from a holistic perspective, including most of the factors that link airspace and airside and affect the overall airport capacity. Furthermore, another contributions of this work is the employment of an approach that uses optimization together simulation, with the objective of obtaining more robust solutions. On one hand, optimization allows to come up with optimal or suboptimal solutions and, on the other hand, simulation allows to take into account other aspects as stochasticity and interactions inside the system. In literature we can find similar works that employ optimization together with simulation techniques, like in (Mujica, 2015) where an evolutionary algorithm together with discrete event simulation was proposed for the improvement of the check-in allocation, or (Arias et al., 2013), where was presented a model for solving the stochastic aircraft recovery problem employing constraint programming together with simulation.

In this paper a methodology for detecting and resolving conflict at airports is presented, it considers the airport from a holistic view, taking into account both airspace and airside components together. In the methodology presented in this paper an optimal or sub-optimal solution is found applying a sliding window approach (Hu and Chen, 2005; Zhan et al., 2010; Furini et al., 2015) together with a meta-heuristic (simulated annealing) (Kirpatrick et al., 1983), after that, the solution provided from the optimization model is tested and validated with the use of a discrete event simulation

model. Using simulation, it is possible to take into account the stochasticity of the system and the interactions between the entities in the system.

The paper is organized with the following structure, in Section 2 the methodology is explained, in Section 3 a scenario is tested, providing preliminary results and in Section 4 some conclusions are drawn and next steps for this research are delineated.

2 Methodology

The methodology presented in this work is constituted by three main steps (Figure 1). The first step aims at modeling the airport taking into account both airspace and airside. In this case the airside was modeled in a "macro" level, where runway, taxiway and terminals were modeled as servers with a specific capacity. The second step consisted in the implementation of an optimization model to obtain a solution for the conflict detection and resolution problem. Finally, in the third step, the solution provided by the optimization model is evaluated by the means of a simulation model in order to test the effectiveness and feasibility of the solution.

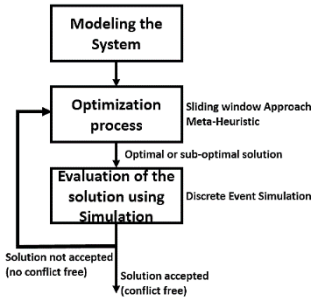


Figure 1. Methodology steps.

2.1 Airport Modeling

One of the main contribution of this work is that it considers both airspace and airside of the airport. Concerning the airspace, landing routes in the TMA were modeled, and separation minima between aircraft as well as speed restrictions were included. Regarding the airside, since the objective did not require a detailed evaluation, runway, taxiway and terminals components, were modeled as servers with a specific capacity. First, it is fair to explain the concept behind airspace and airside conflicts. In this framework, it was assumed that any violation of separation minima between aircraft along the airspace routes and at the merging point was considered as a conflict. Values about separation minima are in accordance with the ICAO standards for separation minima due wake vortex turbulence, they are based on the aircraft type which could be light, medium or heavy (see Table 1). Concerning the airside, conflicts were detected when the capacity of runway, taxiway and terminal was exceeded. It is clear that, the objective of detecting and resolving conflict in the airspace and in the airside lead to have a smooth flow of aircraft in the airspace and a balanced load on the airside, which is the main scope of this work. In this work, the case of Paris Charles de Gaulle Airport was considered. Regarding the airspace, standard approach routes (STAR) and final approach segment were modeled. In total there are four different routes coming from different entry points, all of them merge at the merging point before the final approach segment. In Figure 2 the airspace routes taken into account in the model are shown.

Table 1. ICAO wake vortex turbulence separation minima.

		Leading Aircraft		
		Heavy	**Medium**	**Light**
Trailing Aircraft	**Heavy**	4	3	3
	Medium	5	4	3
	Light	6	4	3

As a preliminary test, in the model there were considered only one of the three terminal and only two runway (one landing and one departing) of the four runways (two landings and two departing) that constitute the airport airside. Concerning the values chosen for the capacity of the airside components, it is intuitive that runways have the value of capacity equal to one since only one aircraft is allowed to cross the runway at a time. Of the three terminal, terminal 2 was chosen for being tested, due to the availability of data concerning inbound and outbound flight to and from this terminal. Terminal 2 is the biggest of the three terminals in Paris CdG airport, it accommodates all the flights of Air France and SkyTeam members. In Table 2 all the characteristics concerning airside components are listed with their respective values.

Table 2. Characteristics of airside components.

Airside Component	Capacity
Runway 26 R/L	1
Taxiway Network	20
Terminal 2	176 (152 considered)

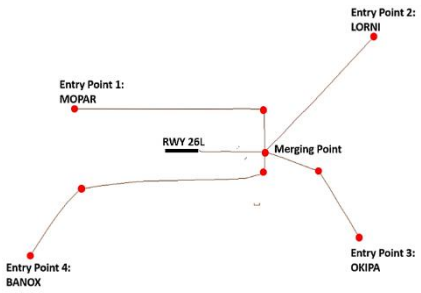


Figure 2. STAR and final approach segment for Rwy 26L.

2.2 Optimization Model

The optimization model proposed to solve the conflict detection and resolution problem is based on (Ma et al., 2015), where a sliding window approach (Hu and Chen, 2005; Zhan et al., 2010; Furini et al., 2015) is used together with a meta-heuristic (simulated annealing) (Kirkpatrick et al., 1983) to solve conflicts in the airspace. The sliding window approach allows to consider an extended time horizon in smaller time frames, dividing the overall problem in sub-problems of smaller size, therefore, decreasing the computational time. Moreover, it allows to treat the problem in a dynamical way, where decisions that are made in each window will affect the decisions to be made in the successive window. The main parameters of the sliding window approach are the size of the window and the size of the shift.

The meta-heuristic used is the simulated annealing (Kirkpatrick et al., 1983), this heuristic is a local search algorithm which is able to escape from local optimum by allowing hill-climb moves in order to find a global optimum.

The main aspect that differentiates this work from the other aforementioned works is that, beside the airspace, airside operations were also included. The new objective becomes the detection and resolution of conflicts both in the airspace and in the airside. The objective of this optimization model is twofold, first it aims at detecting and resolving conflicts in the airspace and also capacity conflicts in the airside, and second is to ensure a smooth flow of aircraft in the airspace and a balanced load on the ground. The decision variables for the problem are: entry time change, entry speed change and pushback time change. The first is the time when aircraft enter the airspace route, the second is the speed that aircraft have when they enter the airspace route and the third is the delay allowed to the aircraft, that are parked at the gate, before they leave the gate and reach the runway for taking off. In Table 3 are shown the values that the decision variables can assume.

Table 3. Value range of the decision variables.

Decision Variable	Value
Entry Time	Between -5 and + 30 min
Entry Speed	Between -10 and + 10
Pushback Time	Between 0 and 5 min

In this context, conflicts in the airspace are detected in the following way: node and link detection. Routes are modeled as a network made by nodes and links, in every node and in every link aircraft are tracked by their "time in" and "time out". If the time interval between "time in" and "time out" overlaps for two or more aircraft then a conflict is detected, the same principle is applied for nodes and links. In Table 4 are described the four routes of the airspace plus the final approach route.

Table 4. Characteristics of the routes modeled in the optimization model.

STAR (Entry point)	Number of nodes (links)
STAR1 (MOPAR)	4 (4)
STAR2 (LORNI)	2 (2)
STAR3 (OXIPA)	2 (2)
STAR4 (BANOX)	4 (4)
Final Approach segment (Merging point)	3 (3)

Concerning airside components as runways, taxiway network and terminal, there were made some assumption about runway occupancy time (for landings and take offs), taxiway occupancy time and turnaround time, they were based on fixed values, in Table 5 these values are listed.

Table 5. Times for airside components in the optimization model.

Airside component	Time
Runway	60 sec landing – 25(H)-30(M)-35(L) sec take off
Taxiway	10 min
Terminal	OffBlockTime-InBlockTime

2.3 Simulation Model

The simulation model is built using a discrete event simulation approach. The employment of this approach allows to take into account the stochasticity of the system and also the interactions inside the system (Banks et al., 2010). For example, values related to runway occupancy time, taxiway occupancy time and turnaround time were modeled following probability distributions, whereas, in the optimization model these values were assumed as deterministic. Another factor that differentiate the simulation approach from the optimization one is the speed profile. In the optimization model the acceleration used is fixed and the time in and time out for each node and link is calculated in a static way based on the length of the link, whereas in the simulation model, speed is regulated using a fixed acceleration that is updated each second. In this way, the speed profile will be more realistic. Moreover, during the descending approach the simulation model does not allow aircraft to fly below a certain speed threshold, indicated as lower bound speed. In the simulation model, although were modeled the same routes, these routes were modeled using more nodes and links. The theory behind this approach is to construct a network of equidistant nodes in order to detect conflict more accurately along the route. In the model nodes are distanced by 5 NM from each other, which is assumed as an acceptable distance to make sure to do not miss any conflict along the route. Based on that,

it is likely to find, under the same conditions, more conflicts in the simulation model than in the optimization model. Due to this network structure, in the simulation model conflicts are detected only on the nodes and not on links. Concerning the detection of conflicts, it is applied the same principle as in the optimization model. In Table 6 the main characteristics of the airspace network are listed.

Table 6. Route network in the simulation model.

STAR (Entry point)	Number of nodes (links)
STAR1 (MOPAR)	11 (11)
STAR2 (LORNI)	7 (7)
STAR3 (OXIPA)	7 (7)
STAR4 (BANOX)	15 (15)
Final Approach segment (Merging point)	3 (3)

The main objective of the simulation model is to test if the solution that comes from the optimization model is feasible also in a more accurate scenario.

Regarding the detection of conflicts on the airside, in the simulation are used the same principles as in the optimization model, with the only difference that times are based on probability distributions and therefore also variability is included, instead, in the optimization model times were based on fixed values, making it more predictable and static.

2.3.1 Validation of the Simulation Model

In order to validate the model, we have conducted a cross validation (Geisser, 1975). This type of validation consists in dividing the set of data in two parts, called set training set and testing respectively, and then using the training set for calibrating the simulation parameter in order to get an outcome that resembles the trailing set data sample. After that, the testing set and the simulation outcome are compared in order to see if the simulation is a good predictive of the data. In this case it was used the first half of the input flight schedule as training set and the second half of the input flight schedule as testing set. Figures 3 and 4 show the daily trend of traffic of the testing set and the simulation outcome after the calibration of the data. In Table 7 are shown the parameters that were used in the simulation model. Figures 5 and 6 show the outcome from the simulation and the testing set. In both sets, Figures 3 and 4 and Figures 5 and 6, it can be seen that the hourly traffic from the simulation and the hourly traffic obtained from the real flight schedule follow the same trend. The mean square error estimator was used to estimate the accuracy of the result obtained from the simulation compared to the real data set.

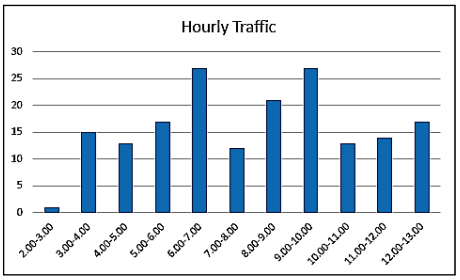


Figure 3. Trend of the daily traffic from the training set.

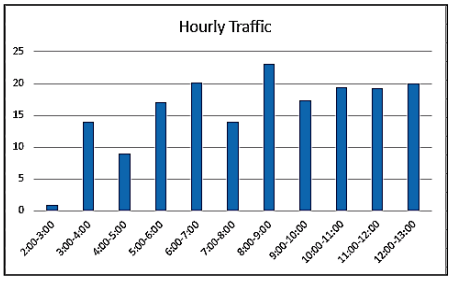


Figure 4. Outcome from the simulation model using the training set as data sample.

In Tables 8 and 9 the values of mean square errors for each hour of traffic are listed, for the training set and testing set, respectively. By observing these values it can be noticed that the ones related to the testing set are not that high which means that the simulation is relatively reliable in predicting the system.

Table 7. Parameters of the airside components in the simulation model.

Airside component	Time
Runway	Triangular(0.5,0.75,1) min
Taxiway network	Triangular(8,10,12) min
Terminal	Triangular(25,35,45) min

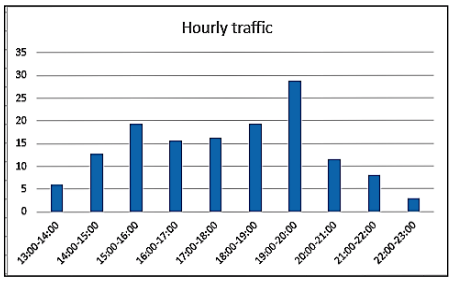


Figure 5. Outcome from the simulation model for the second half of the data sample (testing set).

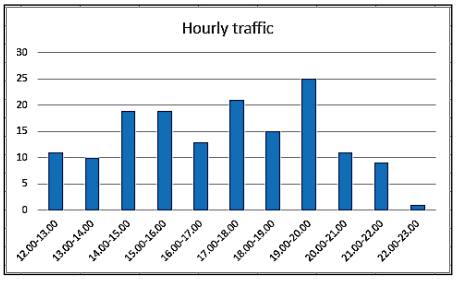


Figure 6. Trend of the daily traffic from the testing set.

Table 8. Mean square error values for training set.

Hour	mean square error
2:00-3:00	0
3:00-4:00	1
4:00-5:00	15.76
5:00-6:00	0.2
6:00-7:00	47.16
7:00-8:00	8.73
8:00-9:00	93.76
9:00-10:00	43
10:00-11:00	29.2
11:00-12:00	10.53
12:00-13:00	9

Table 9. Mean square error values for testing set.

Hour	mean square error
13:00-14:00	16
14:00-15:00	38.6
15:00-16:00	0.66
16:00-17:00	7.6
17:00-18:00	23.8
18:00-19:00	20.2
19:00-20:00	16.5
20:00-21:00	0.7
21:00-22:00	9.3
22:00-23:00	4

3 Scenario and Results

In order to test the goodness of the methodology there have been conducted a series of preliminary tests primarily to tune the parameters of the optimization model in the specific the parameters of the sliding window approach and of the simulated annealing meta-heuristic. Values related to the sliding window parameters are listed in Table 10.

Table 10. Sliding window parameters.

Parameter	Value
Window duration	2Hrs
Shift	30 min

Concerning the experiments, one scenario was tested based on a flight schedule related to a specific day. Once the optimized solution is ready it will be tested with the use of the simulation model developed and discussed in Section 2.3.

3.1 Scenario

In the flight schedule that was used, three different typology of flights were identified: arrivals, departures and arrivals and departures. The first type means that aircraft will arrive and stay at the gate for the whole day without departing again to another destination, in the flight schedule you can find most of them during the middle of the day and in the evening. The second type, is departure flight which means that those aircraft are already parked in one of the gate and they will depart to a destination, usually you can find those type of flights during the morning. Finally, the third type, arrivals and departures, are those flights that arrive from an origin, they park at the gate and then they depart to another destination, these flights represent the majority of the flights in the schedule. In Table 11 the structure of the flight schedule, according to the flight type, is showed.

Table 11. Flight schedule structure for the scenario.

Scenario	
Arrivals	149
Departures	48
Arrivals and departures	91

As it can be seen from the figures above, in the first schedule there are 3 peaks: 7.00-8.00, 10.00-11.00 and 20.00-21.00, with 31, 24 and 25 air traffic movements, respectively.

3.2 Results from the Optimization Model

Table 12 summarize the results obtained running the optimization model before and after the implementation of the simulated annealing meta-heuristic. Looking at the table above, it can be noticed that the optimization model is able to reach a conflict free situation in 120 sec., when the initial solution without optimization registered in total 307 conflicts, 121 on nodes, 144 on links and 42 on the runway. Moreover, it can be noticed that taxiway and terminals are not affected by conflicts both before and after the optimization process, proving that under the given traffic, the capacity of the two components is able to handle this traffic without incurring in congestion problems.

Table 12. Results before and after optimization process.

	before opt	after opt
Computational time	101.883 sec	120.18 sec
Total objective	307	0
Node conflicts	121	0
Link conflicts	144	0
Runway conflicts	42	0
Taxiway conflicts	0	0
Terminal conflicts	0	0

3.3 Results from the Simulation Model

After running the optimization model and obtaining an optimal solution, this solution has been tested by means

of a discrete event simulation model. It was simulated the whole day and there were run 30 replication. Tables 13 and 14 show the results obtained by the simulation model. For simplicity, we have named as the conflict detected on the nodes in the airspace routes were named "airspace conflicts". Moreover, in order to have a better idea of how many aircraft are involved in the conflicts, the number of aircraft affected by at least one conflict in the airspace were collected.

Table 13. Results from simulation model before the optimization process.

	Min	Avg	Max	St. dev
Aircraft conflicts	57	57	57	0
Airspace conflicts	256	256	256	0
Runway In conflicts	86	92	98	1.139
Runway Out conflicts	1	1.55	3	0.4255
Taxiway conflict	0	0	0	0
Terminal conflicts	0	0	0	0

Table 14. Results from simulation model after the optimization process**.**

	Min	Avg	Max	St. dev
Aircraft conflicts	38	38	38	0
Airspace conflicts	180	180	180	0
Runway In conflicts	49	51.3	55	0.7385
Runway Out conflicts	1	1.41	2	0.3272
Taxiway conflict	0	0	0	0
Terminal conflicts	0	0	0	0

In the simulation model it is likely to find more conflicts in the airspace than in the optimization model, due to its different route structure that provides more nodes and links than the route structure of the optimization model. Another distinction was made between "runway in" and "runway out", which refers to runway used for landings and runway used for departures, respectively. Looking at the results, we can see that there are still conflicts, they are mainly concentrated in the airspace were we have 180 conflicts on nodes occurred to 38 aircraft. No source of variability affect these results because there is not any source of stochasticity in the values related to the airspace. We have in average 51.53 conflicts for runway in and 1.41 conflicts for runway out, while no conflicts are detected for taxiway and terminal. It is noticeable that, even though after the optimization process in the simulation model there are still conflicts, they have decreased sensibly, compared with the scenario without optimization process. It proves that the solution provided by the optimization process is able to reduce the number of conflicts, but it is not enough to achieve a conflict free solution.

4 Conclusions and Future Work

In this paper, a methodology for detecting and resolving conflicts at airports is presented. The methodology consists in the implementation of optimization together with discrete event simulation techniques in order to come up with more robust solutions. The optimization model was solved using a sliding window approach and the simulated annealing meta-heuristic. With the use of a discrete event simulation model, the methodology aimed at evaluating the solution given from the optimization model, in a real and more accurate environment. In this work, Paris Charles de Gaulle Airport was taken as a case study and one scenario was tested based on the flight schedule of a specific day.

From the results, we found that the optimization was able to find an optimal (conflict free) solution. When the solution was tested using the simulation model it was found that, although conflicts were sensibly decreased compared to the non optimized scenario, there were still a lot of conflicts both in the airspace and on the runway. From this results it is possible to conclude that the solution from the optimization model was not feasible, and therefore, the optimization model needs further refinements in order to produce a more robust ad feasible solution.

Next steps for this research are in accordance with the results, therefore, the optimization model needs to be refined in order to be more accurate. Furthermore, the airside can be modeled in a more detail, including taxiway routes and gate assignment.

Acknowledgements

The authors would like to thank the Aviation Academy of the Amsterdam University of Applied Sciences and Ecole Nationale de l'Aviation Civile for the support to perform this study.

References

P. Arias, D. Guimarans, G. Boosten, M. Mujica. A methodology combining optimization and simulation for real applications of the Stochastic Aircraft Recovery Problem. *In Conference proceedings of the EUROSIM13. Cardiff, U.K, 2013*. doi: 10.1109/EUROSIM.2013.55.

H. Balakrishnan, B. Chandran. Algorithms for scheduling runway operations under constrained position shifting. *Operation Research*, 58(6): 1650-1665, 2010.

J. Banks, J.S. Carson, B. Nelson, D.M. Nicol. Discrete-Even system Simulation. 5th ed. Upper Saddle River, NJ: Pearson. 2010.

J.E. Beasley, M. Krishnamoorthy, Y.M. Sharaiha, D. Abramson. Scheduling aircraft landings: the static case. *Transportation Science*, 34(2): 180-197, 2000.

J.E. Beasley, J. Sonander, P. Havelock. Scheduling aircraft landings at London heathrow using a population heuristic. *Journal of the Operational Research Society*, 52(5): 483-493, 2001. doi: 10.1057/palgrave.jors.2601129.

A. Bolat. Procedures for providing robust gate assignments for arriving aircraft. *European Journal of Operational Research*, 120(1): 63-80, 2000. doi: 10.1016/S0377-2217(98)00375-0.

U. Dorndorf, A. Drexl, Y. Nikulin, E. Pesch. Flight gate scheduling: State-of-the-art and recent developments. *Omega*, 35(3): 326-334, 2007. doi: 10.1016/j.omega.2005.07.001.

Eurocontrol, Industry monitor. Issue N180-181-182. 2016.

F. Furini, M.F. Kidd, C.A. Persiani, P. Toth. Improved rolling horizon approaches to the aircraft sequencing problem. *Journal of Scheduling*, 18(5): 435-447, 2015. doi: 10.1007/s10951-014-0415-8.

S. Geisser. The predictive sample reuse method with aplications. *J. Amer. Statist. Ass*., 70(350): 315-334, 1975.

X. Hu, W. Chen. Receding Horizon Control for Aircraft Arrival Sequencing and Scheduling. *IEEE Transactions On Intelligent Transportation Systems*, 6(2): 189-197, 2005. doi: 10.1109/TITS.2005.848365

H. Khadilkar, H. Balakrishnan. Network Congestion Control of Airport Surface Operations. *Journal of Guidance, Control, and Dynamics*, 37(3): 933-940, 2014.

S.H. Kim, E. Feron. Impact of gate assignment on gate-holding departure control strategies. *In Conference proceedings of the Digital Avionics Systems Conference (DASC), Williamsburg, (Virginia, USA), 2012*. doi: 10.1109/DASC.2012.6382350.

S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598): 671-680, 1983. doi: 10.1126/science.220.4598.671.

J. Ma. Aircraft merging and sequencing problems in TMA. Master Thesis, Enac, 2015.

A. Michelin, M. Idan, J.L. Speyer. Merging of air traffic flows. *In Conference proceedings of the AIAA Guidance, Navigation, and Control Conference. Chicago, Illinois, 10-13 August 2009*.

J. Montoya, Z. Woord, S. Rathinam, W. Malik. A Mixed Integer Linear Program for Solving a Multiple Route Taxi Scheduling Problem. *In Conference proceedings of the AIAA Guidance, Navigation, and Control Conference*. pages 1-15, Toronto, (Ontario, Canada), August 2-5 2010.

M. Mujica. Check-in allocation improvements through the use of a simulation-optimization approach. *Transportation Research Part A*, 77: 320-335, 2015. doi: 10.1016/j.tra.2015.04.016.

M.E. Narciso, M.A. Piera. Robust gate assignment procedures from an airport management perspective. *Omega*. 50: 82-95, 2015. doi: 10.1016/j.omega.2014.06.003.

N. Pujet, B. Declaire, E. Feron. Input-output modeling and control of the departure process of congested airports. *In Conference proceedings of the Guidance, Navigation, and Control Conference and Exhibit*, pages 1835-1852, Portland, (Oregon, USA), August 9-11 1999. doi: 10.2514/6.1999-4299.

S. Rathinam, Z. Wood, B. Sridhar, Y. Jung. A Generalized Dynamic Programming Approach for a Departure Scheduling Problem. *In Conference proceedings of the AIAA Guidance, Navigation, and Control Conference*, pages 1-12, Chicago, (Illinois, USA), August 10-13 2009.

M. Sandberg, I. Simaiakis, H. Balakrishnan, T.G. Reynolds, R.J. Hansman. A Decision Support Tool for the Pushback Rate Control of Airport Departures. *IEEE Transactions on Human-Machine Systems*, 44(3): 416-421, 2014. doi: 10.1109/THMS.2014.2305906.

I. Simaiakis, H. Khadilkar, H. Balakrishnan, T.G. Reynolds, R.J. Hansman. Demonstration of reduced airport congestion through pushback rate control. *Transportation Research Part A: Policy and Practice*, 66: 251-267, 2014. doi: 10.1016/j.tra.2014.05.014.

I. Simaiakis, H. Balakrishnan. A Queuing Model of the Airport Departure Process. *Transportation Science*, 50(1): 94-109, 2015.

Z. Zhan, J. Zhang, Y. Li, O. Liu, S.K. Kwok, W.H. Ip, O. Kaynak. An Efficient Ant Colony System Based on Receding Horizon Control for the Aircraft Arrival Sequencing and Scheduling Problem. *IEEE Transactions on Intelligent Transportation System*, 11(2): 399-412, 2010. doi: 10.1109/TITS.2010.2044793.

C.A. Zuniga, D. Delahaye, M.A. Piera. Integrating and Sequencing Flows in Terminal Maneuvering Area by Evolutionary Algorithms. *In Conference proceedings of the DASC 2011, 30th IEEE/AIAA Digital Avionics Systems Conference, Seattle, United States, 2011*. doi: 10.1109/DASC.2011.6095980.

C.A. Zuniga, M.A. Piera, S. Ruiz, I. Del Pozo. A CD\&CR causal model based on path shortening/path stretching techniques. *Transportation Research Part C*, 33: 238-256, 2013.

Performance Evaluation of Alternative Traffic Signal Control Schemes for an Arterial Network by DES Approach-Overview

Jennie Lioris[1] Pravin Varaiya[2] Alexander Kurzhanskiy[3]

[1]ENPC, France, jennie.lioris@enpc.fr
[2]California PATH, University of California, Berkeley USA, varaiya@berkeley.edu
[3]California PATH, University of California, Berkeley USA, akurzhan@berkeley.edu

Abstract

Evaluation aspects of alternative traffic signal control strategies for an arterial network are studied. The traffic evolution of a signalized road network is modelled as a Store and Forward (SF) network of queues. The system state is the vector of all queue lengths at all intersections. The signal control at any time permits certain simultaneous turn movements at each intersection at pre-specified saturation rates. Two control categories, open loop and traffic-responsive policies are compared under fixed and time-varying demand. The behaviour of the underlying queuing network model manifesting asynchronous nature over time while involving concurrence is modelled according to an event-driven approach virtually reproduced by discrete event simulations. Exploration of the implementation outputs results a pertinent mathematical framework for traffic movement, analysis and signal control design. Subsequently, various metric measurements such as queue bounds, delays, trajectory travel times quantify the actual policy. Moreover, aggregate behaviour as in a macroscopic queuing model is also prompted. Experiments are performed using real data for a section of the Huntington-Colorado arterial adjacent to the I-210 freeway in Los Angeles. Lastly, the meso-micro simulation issues resulting from the employed decision tool, PointQ, are compared with microsimulation and mesosimulation forms of other traffic simulation programs.

Keywords: traffic responsive signal, adaptive control, pre-timed control, max-pressure practical policy, discrete event simulation

1 Introduction

The *management* of an arterial traffic network is considered. Currently open loop plans are frequently employed often associated with optimised offsets aiming to create green waves in order to minimise trajectory delays. (Muralidharan et al., 2015) studies the traffic dynamics in a network of signalised intersections. It is shown that when the control can accommodate the demand then the network state converges towards a periodic orbit while any effects of the network initial state disappears. Adaptive controls are expected to improve the network performance since they the current network state is taken into consideration in real time. (Varaiya, 2013) studies a traffic-responsive "Max-Pressure" traffic control, (Mirchandani and Head, 2001) proposes an adaptive control predicting demand patterns and queues to compute timings to minimize average delay. (Aboudolas et al., 2009) suggests an optimal formulation designing a feedback policy. Research studies (Gomes et al., 2008) characterise the behaviour of the cell transmission model of a freeway divided into N cells each with one on-ramp and off-ramp. It is shown that ramp metering eliminates wastefulness of freeway resources.

The present work, appraises the performance of versions of the adaptive Max-Pressure algorithm under unpredicted demand fluctuation. In particular, feedback signal control designs and their related effectiveness are presented and analysed when applied to a network while they are also compared with open loop schemes. Queueing models are employed when designing closed loop signal control plans evaluated by queue based criteria such as queue delays. Thus, traffic evolution is modelled as a controlled store and forward (SF) queuing system. Identified vehicles arrive in iid (independent, identically distributed) streams at entry links, travel along non-saturated (internal) links, join appropriate queues and leave the network upon reaching exit links. At each time and at each intersection, a set of simultaneously compatible movements or *phases* is actuated. Vehicles are discharged at a service rate determined by the *phase* saturation flow rate. When finite internal link capacities are considered, the vertical point queues become horizontal in the sense that interfere in the link vehicle storage capacity and the related link travel times.

A separate queue is considered for each turn movement at each intersection.

Aiming at evaluating the network performance under different control policies a decision making tool is necessary in order to virtually reproduce the considered structure (intersection node and links, vehicle movements and the related control plans) under multiple traffic conditions for both closed and open loop actuation plans. Thus, measurements of various metrics such as delays, travelled times, vehicle queues etc. will be able to quantified according to the employed strategy.

When considering "driver-behaviour", differential

equations are required, emulating "car-following" and "lane-changing" aspects. Microsimulation models are necessary for which queue sizes are not state variables and saturation flow rates are not input data. Instead, they are derived from the simulation analysis. Consequently, it is not possible to relate delays to timing schemes and these models are unsuitable for traffic control conception.

Macroscopic simulators often based on the cell transmission model (CTM) (Lo, 2001) represent traffic flow as a fluid. Spatial density is required as a state variable which is hard to measure. Furthermore, modelling turns, shared lanes, queues or introducing sensor behaviour for actuated signal control, under such approach is rather a hard work.

A made-to-measure micro-meso simulation decision making tool called *PointQ* maintaining the identity of each single vehicle while ignores the vehicle interaction is introduced. It has minimal data requirements and receives saturation flow rates as explicit input values. The PointQ decision tool is developed according to discrete event approach (Baccelli et al., 1992) in order to accurately reproduce the evolution of the asynchronous system while it is appropriate for modelling open and closed loop timing control schemes.

The rest of the paper is organised as follows. Section 2 presents the problem formulation and briefly recalls the utilised control schemes. Section 5 introduces PointQ and reasons the employed model approach. Section 5 and Section 6 discuss the performed experiences. Finally, Section 7 compares PointQ with other micro and mesosimulation modes.

2 Traffic Regulation: Stage selection

The simultaneously compatible movements of an intersection are represented by a binary matrix U, the *intersection stage* of which the (i,j) entry equals one if the corresponding phase is actuated, zero otherwise. Let $\mathscr{U}$ be the set of admissible *stages* of a given intersection and $\gamma(l,m)$ the turning ratio of phase (l,m), expressed as the probability of a vehicle to choose as destination link m when joining link l. The optimisation horizon is divided into intervals or cycles of fixed width, each one comprising of T periods. Within each cycle, there exist $T-L$ available planning periods where $L<T$ represents the idle time corresponding to pedestrian movements, amber lights, etc. Let q be the array of which the (i,j) entry is the length of queue related to phase (i,j). The system state at time $t, X(t)$ is defined by $X(t)=q(t)$. A control *stabilises* the network, if the time-average of every mean queue length is bounded. At a given time *stage* $u(t)=U, U\in\mathscr{U}$ and $\lambda_{u(t)}$ cycle proportion have to be decided such that:

- $u(t)$ *stabilises* $X(t)$
- if $\tilde{c}(l,m)$ denotes the service rate of phase (l,m) and f_l represents the vehicle flow in link l, then the following stability condition has to be verified,

$$\tilde{c}(l,m) > f_l\gamma(l,m), \tag{1}$$

- $\sum_{u\in\mathscr{U}} \lambda_u T + L \leq T$.

3 Signal Control Schemes

A brief description of the utilised traffic control algorithms is now presented. The related theory is explicitly developed and analysed in (Varaiya, 2013).

3.1 Pre-timed network control

A *fixed-time* control (FT) is a periodic sequence, $\{\lambda_U, U\in\mathscr{U}\}$, actuating each *stage* $u(t)=U^i$, $U^i\in\mathscr{U}$ for a fixed duration $\lambda_{U^i}T$ within every cycle of T periods.

3.2 Max-Pressure Practical (MPract)

Max-Pressure is a distributed policy selecting a stage to actuate as a function of the upstream and downstream queue lengths. The pressure $w(q(t),U)$ exerted by *stage* $U\in\mathscr{U}$, is defined by

$$w(q(t),U) = \sum_{(l,m)} \varsigma(l,m)(t) S\circ U(l,m)(t) \tag{2}$$

where

$$\varsigma(l,m)(t) = \begin{cases} q_{(l,m)}(t) - \sum\limits_{p\in\mathscr{O}(m)} \gamma_{(m,p)} q_{(m,p)}(t), \\ \qquad\qquad\qquad \text{if } q_{(l,m)}(t)>0, \\ 0, \text{ otherwise.} \end{cases} \tag{3}$$

At time t, *Max-Pressure* control, selects to actuate the *stage* exerting the higher pressure to the network,

$$U^*(q)(t) = \operatorname{argmax}\{w(q(t),U), U\in\mathscr{U}\}, MP\ stage \tag{4}$$

The MPract algorithm applies the new selected MP stage if significantly larger pressure w,

$$\max_U w(U,q(t)) \geq (1+\eta) w(U^*,q(t)). \tag{5}$$

Parameter η is related to the desired degree of stage switches.

4 Modelling and Simulation Overview

4.1 An event-driven approach

Traffic control constitutes an asynchronous, complex structure where uncertainty and concurrence are naturally inherent. Many theoretical questions related to which models and methods are best to utilise for evaluating the network performance exist. However, one observes that there are queue-based models and car-following models. To our knowledge, all signal control algorithms use queue-based models. Since, we are concerned by signal control designs, a queue-based approach is appropriate to the needs of the study. Queueing theory is intended to be *descriptive*, given a model and control policies, after analysis, verification issues examine whether the desired objectives are attained and (potentially) performance is obtained.

A Discrete Event System (DES) is a dynamical system the behaviour of which is ruled by occurrences of different types of events over time rather than fixed time steps. Although time evolves within two consecutive events, the sole responsible for state transitions is the event realisation. Differential equations (developed for the analysis of time-driven systems) form no longer an adequate setting. Simulation means consist a reliable way to describe the DES dynamics.

For the study of the arterial management a mesoscopic-microscopic discrete event decision tool, "PointQ", is developed. Vehicle identities are preserved but driver-interaction is intentionally ignored. Vehicle routes are observed and travel times are measured. Moreover, PointQ requires similar model parameters as macroscopic approaches (network geometry, demand and signal control). Since elementary calibration is required based on commonly available field data, PointQ can efficiently evaluate the influence of traffic control algorithms to the network. However, PointQ is inadequate for studies related to driver behaviour effects or specific network geometry.

In a DES approach the system evolution is represented as a chronological sequence of events of the form $\{\ldots, s_i, e_i, s_{i+1}, e_{i+1}, \ldots\}$, where s_i is the system state at time t_i and e_i is an event occurring at time t_i marking changes to the system bringing it to state s_{i+1} and so forth. It is assumed that the system is *deterministic* in the sense the state resulting from an event realisation is *unique.* PointQ model involves events on vehicle arrivals, departures and signal actuation.

4.2 PointQ design

The entire structure is split into two independent but also closely interacting parts according to the task nature. The mechanical part virtually represents the system entity interactions. It receives tow types of entries:

- input data such as network geometry (link capacity, speed limit or mean travel time, turn pocket capacity, phase saturation flow rates which can either be measured or estimated during implementations), initial traffic state, sensors etc.
- controls ruling the system.

On the other hand, the real time management comprised of all the decision algorithms required by the mechanical part involving signal controls, vehicle arrival/departure decisions, demand patterns, routing algorithms, queue estimation models etc.

5 Case Study-Data description

A section of the Huntington-Colorado arterial near the 1-210 freeway in Los-Angeles, comprised of 16 signalised intersections, 76 links and 179 turn movements, is considered. Figures 2 and 3 illustrate the network map and its abstraction as a directed graph. Stochastic (Poisson) external arrivals are employed, generating approximatively $14,500$ vehicles per hour. Utilisation of an identical demand contributes to the accuracy of measurements related to the influence of each control scheme on the network. Vehicle routing is based on turning probabilities. The Fixed-Time plan and the turn ratio values are provided by the local traffic agency. The cycle T is of 120 seconds (with some exceptions at two nodes where the cycle is of 90 and 145 secs). The idle duration corresponding to each cycle is between 0 and 2 seconds. Internal links are of finite vehicle storage capacity. Stochastic travel times based on the free flow speed and the current link state are considered. Time granularity is taken equal to 0.1 seconds while the network evolution is reproduced over a period of 3 hours.

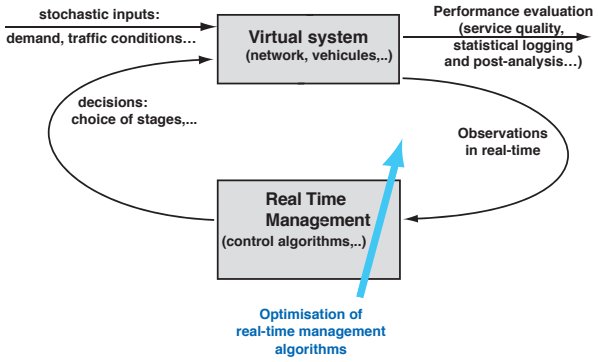


Figure 1. The simulator in two parts.

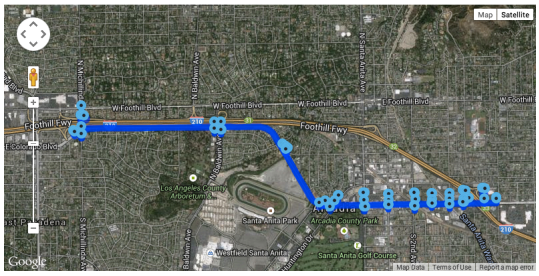


Figure 2. Huntington-Colorado site map.

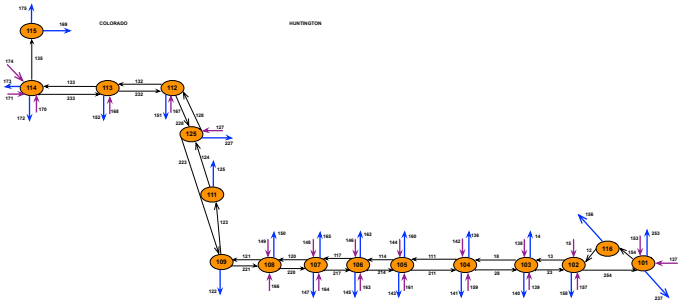


Figure 3. Directed Network graph.

6 From theory to applications

In what follows, we focus on the network performance evaluation for both timing plans, FT Offset and MPract. Metrics on queue lengths, travel times and delays are investigated for two demand patterns, the baseline demand provided by the data and a time-varying one.

6.1 System Stability

Figure 4 plots the network state evolution for both the Pre-timed and MPract signal controls. The theoretically expected demand accommodation is also experimentally verified. Moreover, one observes that the feedback plan (blue curve) maintains lower queue values.

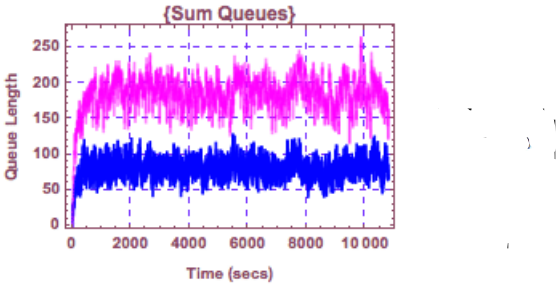


Figure 4. Sum network queues: MPract 8-blue curve, FT Offs-purple curve.

6.2 Trajectory Delay Measurement

For each realised trajectory (sequence of entry, internal and exit links) delays faced by all vehicles followed the related path are measured. The corresponding distribution is computed and the CDF function is represented in Figure 5. MPract (purple curve) reduces delays almost four times in comparison with the fixed time plan (green plot).

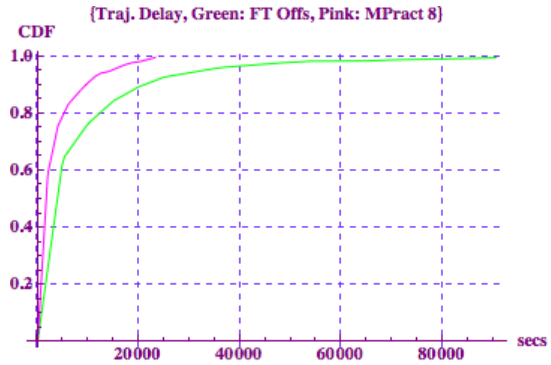


Figure 5. CDF Trajectory Delays: MPract (purple), FT-Offs (green).

6.3 Queue Delay Measurement

Delays on distinct queues on link 114 (incoming link at node 106, Huntington region) are now measured.

Three phases are associated with link 114. Figures 8, 9 and 10 depict the evolution of cumulative delay values for each phase according to the Pre-timed and MPract policies. Observe that for phase $(114,145)$ the pre-timed scheme implies lower delays. Mainly, this is due to the fact that $q(114,145)$ and $q(114,145)$ head vehicles towards exit links. More precisely, phases $(114,145)$ and $(217,214)$ are simultaneously actuated by stage 1. Similarly phases $(114,117)$ and $(217,214)$ are actuated by a concurrent stage 2. At any decision time the MPract stage is the one exerting the higher pressure. According to equation 2 and since no output queues are associated with the exit links 145, 162, the pressure exerted by stage 2 is determined by the queue lengths of the related phases. Taking into consideration the flows and queue demand on link 114, stage 2 often exerts higher pressure regarding stage 1. Thus, it receives increased green time duration. Figures 6, 7 plot the evolution of queues $q(114,145)$ and $q(114,117)$ for both policies. Lower queues result under MPract for phases of stage 2. The evolution of cumulative delay values for each phase of link 114 is represented in Figures 8, 9 and 10 (saturation flow rates remain unchanged for both policies).

Tables 1, 2 resume the mean time spent by vehicles in four queues and the average vehicle sojourn time in all queues respectively. Obviously, MPract appears more refined although for phase$(170,173)$ FT-Offset implies smaller delays. Table 3 presents the total travel time between three entry-exit links.

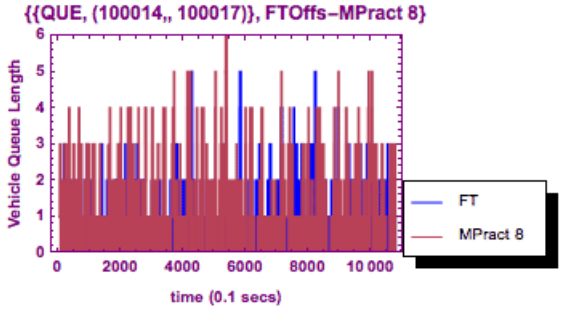


Figure 6. Evolution q(114,117), FT Offs-MPract 8.

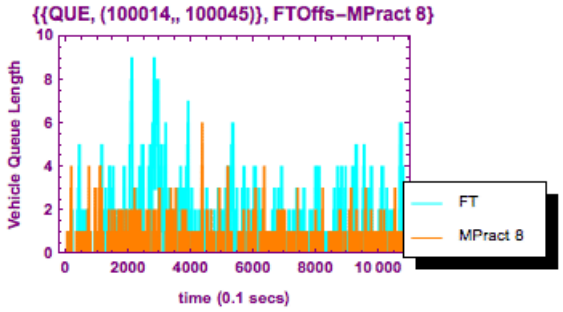


Figure 7. Evolution q(114,145), FT Offs-MPract 8.

6.4 Varying Traffic Conditions

The network stability is now examined under time-varying traffic intensity. During a given period, demand gradu-

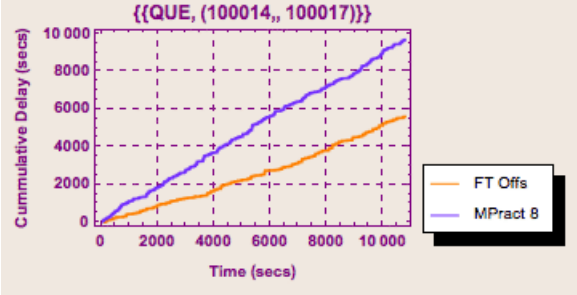


Figure 8. Cumulative Delay q(114,117), FT Offs-MPract 8.

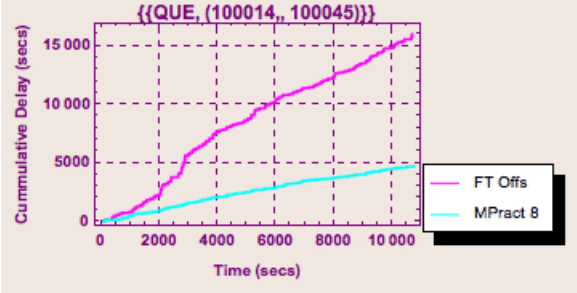


Figure 9. Cumulative Delay q(114,145), FT Offs-MPract 8.

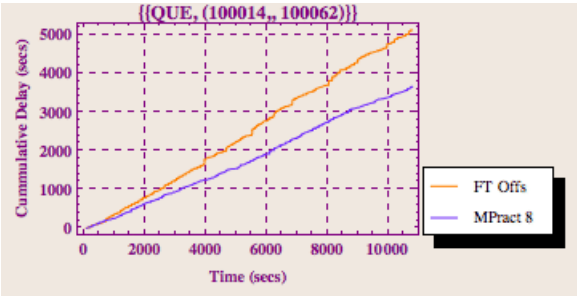


Figure 10. Cumulative Delay q(114,162), FT Offs-MPract 8.

Table 1. Mean Time spent by vehicles in queue.

Que ID		Mean Veh. Sojourn Time (secs)	Mean Veh. Sojourn Time (secs)
		MPract	FT Offs
164	1020	49.96	79.72
167	228	12.16	51.32
117	120	10.72	33.35
170	173	78.46	49.61

Table 2. Average Mean Time spent by vehicles in all queues.

AVERAGE MEAN VEH SOJOURN TIME IN QUEUES	MPract (secs)	FT Offs (secs)
	12.05	22.32

Table 3. Travel Time Entry-Exit Link.

Entry link	Exit link	Travel ttime (secs)	Travel time (secs)
		MP	FT Offs
127	145	204.4	230.8
164	173	364.7	462.2
138	1069	395.2	519.6

ally increases and potentially temporary congested conditions may result (representing peak hours or unpredicted demand variation). Thus, within period $[0, 36,000]$ external demand increases every two hours by a factor c_1 equal to $1.05, 1.15, 1.3, 1.5$. For the following two hours, that is from $t = 36000.1$ to $t = 43,200$ seconds, demand decreases every half an hour by a factor c_2 taking progressively values $1.3, 1.15, 1.05, 1$.

Figure 11 illustrates the evolution of all the network queues when the pre-timed actuation durations are employed. While demand remains inferior to $1.15d$ (d is the initial demand level), that is while time $t < 14,000$ secs the system remains stable. Congestion spreads during period $[14,400, 36,800]$ when the current demand intensity values $1.15d$, $1.3d$, $1.5d$. During this time a significant portion of the network links become congested, strongly increasing the number of vehicles in the network. Obviously, this Fixed-Time plan cannot accommodate the new demand intensity. The resulting FT behaviour is theoretically expected from the stability condition presented in §2. When the demand level decreases link saturation progressively diminishes (from $t = 36,000$ to $43,200$ seconds).

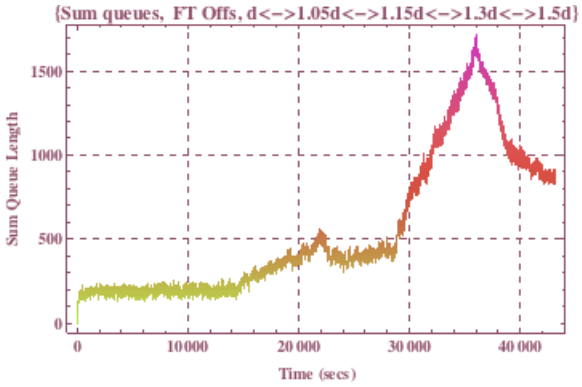


Figure 11. Evolution sum queues, varying demand level, FT Offset.

In contrast, the network behaviour differs when a MPract policy defines signal plans. Figure 12 depicts the sum of the network queues for the $c_i d$, $i = 1, 2$ demand intensities. Clearly, the feedback policy prohibits link saturation. The sum of all the network queues rises during period $[21,600, 36,000]$ when a 30% and 50% increase of the initial demand takes place but still the network remains stable.

Figures 13 and 14 plot the aggregate entry (blue

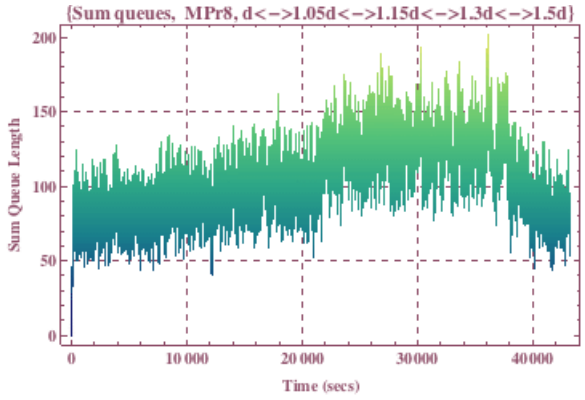


Figure 12. Evolution sum queues, varying demand intensity, MPract 8.

curve) end exit flows (red, purple curves) during period $[29,000, 31,000]$ for FT and MPract policies respectively. Since, the pre-timed control cannot accommodate $c_1 d$, demand for $c_1 = 1.3, 1.5$ the number of vehicles exiting the network drops below the external arrivals. This phenomenon disappears under MPract.

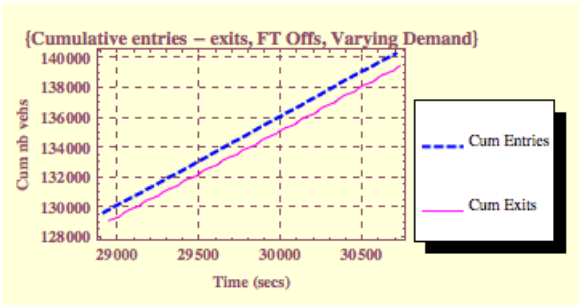


Figure 13. Aggregate entries-exits, varying demand, FT Offs.

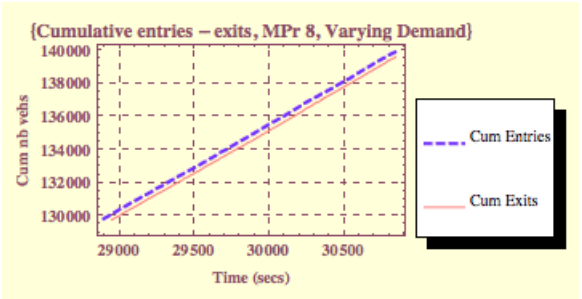


Figure 14. Aggregate entries-exits, varying demand, MPract 8.

7 PointQ versus AIMSUN Network Performance

As previously discussed, PointQ is a micro-meso decision tool, approaching a SF network queue model by a discrete event technique.

AIMSUN is a vehicle traffic software offering mesoscopic, microscopic and hybrid simulation approaches.

Aiming at a comparison of the two simulation tools, the section of Huntington-Colorado arterial is modelled in AIMSUN. Both approaches employ stochastic demand of intensity d (baseline demand), Pre-timed signal plan as governing control and consider the turn ratio values and link free flow speeds as provided by the data. Microscopic and Mesoscopic AIMSUN simulations are performed for a three hours duration. The state of intersection 101 (first node at the Huntington area) is investigated according to PointQ and AIMSUN implementations.

Five controlled movements exist at node 101 corresponding to queues $q(137,154)$, $q(153,154)$, $q(153,237)$, $q(254,237)$ and $q(254,253)$. We focus on the behavior of two representative movements. Phase $(137,154)$ brings vehicles into the network from the entry link 137 and head them towards the internal link 154 while phase $(254,237)$ moves vehicles from the internal link 254 towards the exit link 237.

7.1 Evolution of phase $(137,154)$

Figure 15 represents the evolution of queue $q(137,154)$ under a PointQ simulation. Figures 16 and 17 illustrate the queue behaviour when micro and meso AIMSUN simulations are performed.

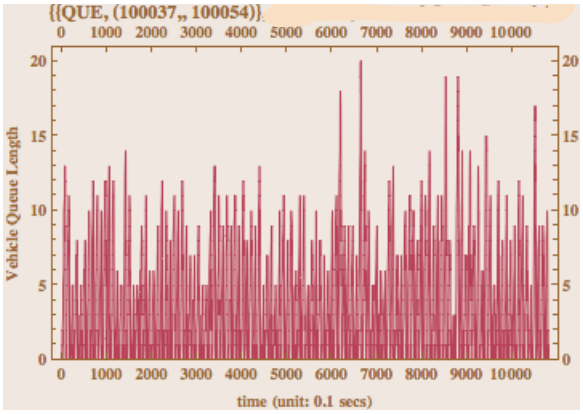


Figure 15. Evolution of q(137,154), PointQ.

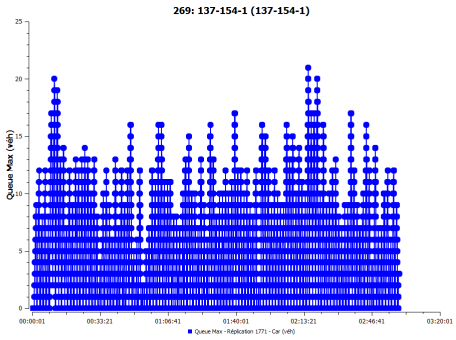


Figure 16. Evolution of q(137,154), AIMSUN Micro.

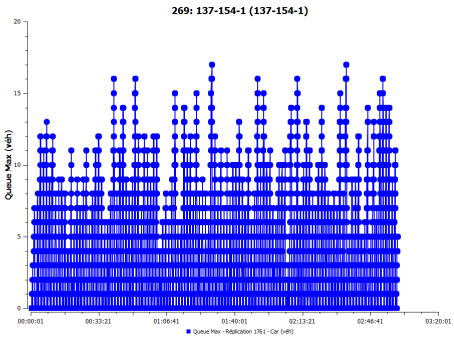


Figure 17. Evolution of q(137,154), AIMSUN Meso.

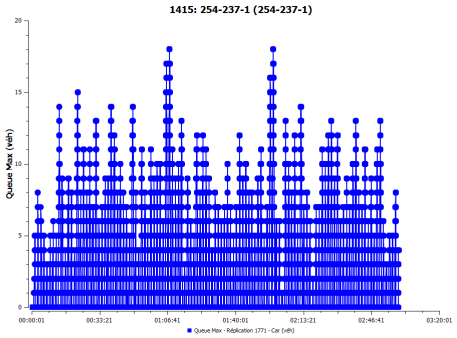


Figure 19. Evolution of q(254,237), AIMSUN Micro.

Although stochastic external demand and link travel times are considered, one observes that the three approaches, PointQ model and AIMSUN micro and meso versions provide close results. Queues resulting from the three simulation modes verify the theoretically resulting stability. Furthermore, queue lengths vary within similar bound values over time.

7.2 Evolution of phase $(254, 237)$

Figure 18 plots the behaviour of queue $q(254, 237)$ when PointQ while Figures 19 and 20 describe the resulting queue state under micro and meso AIMSUN approaches.

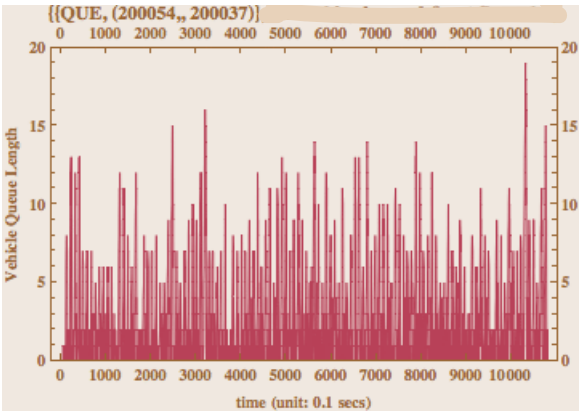


Figure 18. Evolution of q(254,237), PointQ.

As in the case of phase $(137, 154)$, queue $q(254, 237)$ shows the same behaviour for both PointQ and AIMSUN models and all modes (micro-meso, micro and meso).

An extended analysis of all the network queues implies that the resulting network state is similar under the two models.

8 Conclusions

Aiming at a further improvement of traffic, new signal schemes are designed, evaluated and potentially optimised before a real time application. The key contribution of this paper is to present the extended options of a microscopic-mesoscopic decision tool, called PointQ destined for an

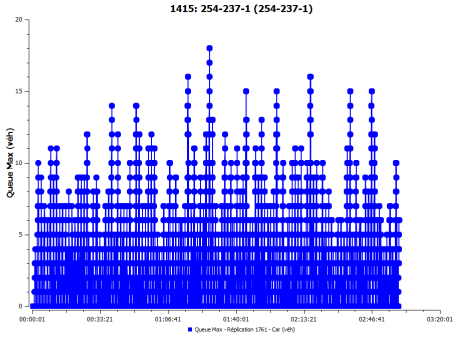


Figure 20. Evolution of q(254,237), AIMSUN Meso.

an ameliorated study of arterial traffic. PointQ relies on the principle of discrete events and models arterials as a Store and Forward queuing network. Thus, queues form state variables of the system. Minimal explicit input information is needed amongst which the saturation flow rates. These values are necessary to signal control development since most feedback algorithms follow queue-based approaches. Experiments are performed under real data for a section of Huntington-Colorado near Los Angeles. Two control policies are employed, Pre-timed and Max-Pressure Practical plans under fixed and time-varying demand intensity. The expected theoretical results concerning the network stability are verified and the network performance is quantified in terms of queue bounds, delay metrics and travel times. Finally, PointQ accuracy is observed in comparison with a simulation program providing both micro and meso simulation modes. Useful directions of future work worthing to be pursuing is the development of queue estimation algorithms, employed in feedback controllers (e.g. versions of MP algorithm) and actuated controls as well. Moreover, introducing additional modes of sensor behaviour the combination of which would improve precision in the computational results (actuated controls).

References

K. Aboudolas, M. Papageorgiou, and E. Kosmatopoulos. Store-and-forward based methods for the signal control problem in large-scale congested urban road networks. *Transportation Research Part C-Emerging Technologies*, 17:163–174, 2009.

R. E Allsop. Delay-minimizing settings for fixed-time traffic signals at a single road junction. *IMA Journal of Applied Mathematics*, 8(2):164–185, 1971.

F. Baccelli, G. Cohen, J.Olsder, and J.P.Quadrat. *Synchronization and Linearity – An Algebra for Discrete Event Systems*. Wiley, 1992. ISBN ISBN 0 471 93609 X.

G. Gomes, R. Horowitz, A. Kurzhanskiy, J. Kwon, and P.Varaiya. Behaviour of the cell transmission model and effectiveness of ramp metering. *Transportation Research Part C: Methodological, 16(4):485-513, August 2008*, 2008.

B.G Heydecker. Objectives, stimulus and feedback in signal control of road traffics. *Journal of Intelligent Transportation Systems*, 8:63–76, 2004.

H. K Lo. A cell-based trafic control formulation: Strategies and benefits of dynamic timing plans. *Transportation Science*, 35(2), 2001.

P. Mirchandani and L. Head. A real-time traffic signal control system: Architecture, algorithms, and analysis. *Transportation Research Part C-Emerging Technologies*, 9(6):415–432, 2001.

A. Muralidharan, R. Pedarsani, and P.Varaiya. Analysis of fixed-time control. *Transportation Research Part B: Methodological, 73:81-90, March 2015*, 2015.

TRB. Highway Capacity Manual. Technical report, Transportation Research Board, 2010.

P. Varaiya. Max pressure control of a network of signalized intersections. *ACM Transactions on Mathematical Software*, 2013.

Formal Verification of Multifunction Vehicle Bus

Lianyi Zhang[1,2] Duzheng Qing[1,2] Lixin Yu[2] Mo Xia[3] Han Zhang[2] Zhiping Li[2]

[1] Science and Technology on Special System Simulation Laboratory, Beijing Simulation Center, China
[2] State Key Laboratory of Intelligent Manufacturing System Technology, Beijing Institude of Electronic System Engineering, China
[3] Software School, Tsinghua University, Beijing, China

Abstract

Multifunction Vehicle Bus (MVB) is a critical component in the Train Communication Network (TCN), which represents a challenging problem for model checking. Although model checking is widely used in circuit and software verification, it is hardly for verification of the MVB or TCN, in terms of modelling of MVB components and making appropriate specification. The study described in this paper aims at evaluating and experimenting the industrial application of verification by model checking, and provides a complete system modelling and specification describing technique. The model of MVB consists of device model, communication model and specification model translated from LSCs, a scenario description. Experiments results with SPIN checking tool illustrate effectiveness of our approach.

Keywords: vehicular communication, protocol verification, model checking, specification

1 Introduction

Multifunction Vehicle Bus (MVB) in Train Communication Network (TCN) is widely used in most of the modern train control techniques of transportation software system. MVBs change roles of master or slaves under mastership transfer protocol in IEC standards (Kirrmann and Zuber, 2001). How to ensure security of an embedded vehicle control software system which implements the function and protocol has become an important issue.

The traditional method to verify MVB uses simulation technique (Jiménez et al., 2006) or test methods (Zhiwu et al., 2008). These approach need semi-finished MVB devices and programs, error-tolerance decode algorithm and samples of transmitted data among devices. However, simulation and test cannot provide completeness verifica-tion.

Model checking (Queille and Sifakis,1982; Clarke et al., 1999) is a method for automatically verifying finite state systems, using an exhaustive search of the state of a system model to determine whether a specification is satisfied or violated, which has been applied in circuit and software verification. Although model checking is widely used, it is hardly for verification of the MVB or TCN, in terms of modelling of MVB components and making appropriate specification. System modelling prefers fewer states, avoiding states space explosion, and specification making requires different properties synthesis with light weight manual work.

This paper presents a modelling approach for MVB based on finite state model checking and a specification generated method. The remainder of this paper is organized as follows. The following section gives background details of MVB, mastership transfer protocol and preliminaries model checking with temporal logic. Section 3 propose a modelling method for system component and communication with process and finite state automata. Section 4 show property generated and modelling method compound with previous system model. Experiment results in Section 5 demonstrate our methods and Section 6 make conclusion.

2 Background

In this section, we first introduce the MVB and the master transfer protocol. Then model checking with temporal logic is reviewed.

2.1 MVB and Master Transfer

2.1.1 Multifunction Vehicle Bus

The on-board train communication system has been widely used for modern railways. The MVB is a component of the TCN which is used in most of the modern train control systems. The TCN has been defined by the IEC(international Electrotechnical Commission); it is the Vehicle Bus specified to connect standard equipment. It provides both the interconnection of programmable equipment pieces amongst themselves and the connection of this equipment with its sensors and actors. It can also be used as a Train Bus in trains which are not separated during normal operation.

The MVB defines two types of devices: Master and Slave. Each Vehicle Bus and Train Bus has one Master node and several Slave ones. The Master sends information, the *Master_Frame*, to a number of Slave devices. The Slave receives information from the Bus and sends information, the *Slave_Frame*, in response to the Master. A Master_Frame and the corresponding Slave_Frame form a telegram. All devices decode the Master_Frame. The addressed source device then replies with its Slave_Frame, which may be received by several other devices.

2.1.2 Mastership Transfer

Since a single Master presents a single point of failure, mastership may be assumed by several Bus Administrators(BAs), one at a time. To increase availability, mastership can be shared by two or more BAs, which both take charge of mastership for the duration of a turn. Mastership is transferred from BusAdmin to BusAdmin within a few milliseconds in case of failure. To exercise redundancy, mastership is transferred every few seconds by a token frame. Consequently, all BAs are organized in a logical ring. A token passing mechanism ensures that only one BusAdmin become Master.

In the IEC 61375-1 international standard, Mastership Transfer describes the protocol which selects a Master form one of several BAs and ensures Mastership Transfer at the end of a turn or upon the occurrence of a failure. A token passing algorithm is defined in the IEC standard to ensure a round-robin access of all BAs to the Bus:

1. after the loss of the Master, staggering of the timeouts ensures that only one of the BAs become Master;
2. a Master exercise mastership for the duration of one turn;
3. after its turn, the Master looks for the next BusAdmin and reads its Device Status, which indicates if this device is a configured BusAdmin;
4. a Master may only pass mastership to a configured and actualized BusAdmin;
5. if the device is not a configured and actualized BusAdmin, the Master looks for the next BusAdmin after the next turn;
6. if the device is a configured and actualized BA, the Master offers mastership to it by sending a Mastership Transfer Request;
7. if the device accepts mastership in its Mastership Transfer Response, or if no answer comes, the Master retires to become a standby Master and monitors the Bus traffic for mastership offer or Bus silence;
8. if the other device rejects mastership, the current Master retains mastership for one more turn, after which the Master tries the next device in its BAs list;
9. a standby BusAdmin becomes Master if it accepts a Mastership Transfer Request or if it detects no Bus activity during a time greater than a defined timeout;

2.2 Model Checking with temporal logic

For model checking system against some specification, we need both system model structure and description of property. Communicating Finite State Machines (CFSMs) are natural models for systems of concurrently running process, especially asynchronous reactive system. The concurrency of process is captured at the semantics level of CFSMs by the interleaving of processes executions. Process exchange messages between each other asynchronously over a set of message channels, which are interpreted at the semantic level as unbounded FIFO message queues. A sender process continues its local execution after sending a message to a channel, and a receiver process is blocked when it tries to receive a message that is not available in the respective channel.

Definition 1 *(Communicating Finite State Machines). A system of communicating finite state machines (CFSM) is a tuple* $\mathscr{S} = (P,V,M,C,succ)$*, where*

- *P is a finite set of* process*. Each process* p_i *is a pair* (S_i, s_0^i) *where* S_i *is a finite set of states of* p_i *and* $s_0^i \in S_i$ *is the initial state. For any two different process* p_i *and* p_j*, we put the restriction that* $S_i \cup S_j =$*, i.e., their sets of states are disjoint.*
- *V is a finite set of* variable*, may be global or local in process.*
- *M is a finite set of* message symbols*.*
- *C is finite of* messages channels*. Each channel is associated with a subset of message symbols* $M' \subset M$ *such that only the messages in* M' *can be exchanged in the buffer. Moreover, for each channel* $c \in C$ *and each message symbol m in the subset of M associated with c, we call* (c,m) *a message type.*
- *succ is a finite set of* local transitions (s,e,s') *where s and* s' *are states of some same process* p_i*, and e is* (g,u)*: g is* guard condition *consist of both boolean formula with or without a message passing event in the form b*!*m or b*?*m such that (1)*$c \in C$ *and (2)*$m \in M$ *can be exchanged in the buffer c; u is updates of variables as* $v' = u(v)$*.*

The semantics of CFSMs is defined using the concepts of configurations and reachability.

Definition 2 *(Configuration). Given a CFSM system* $\mathscr{S} = (P,V,M,C,succ)$*, a* configuration*(or* global state*) of the system is a tuple* $(s^1,...,s^{|P|},v,q^1,...,q^{|C|})$ *such that*

- *each* s^i *is a state of the process* p_i*, and*
- *each* q^i *is a queue of messages exchangeable in the channel* c_i*, and*
- *v is a valuation of all variables.*

Consider two configuration s_1 *and* s_2*, let* $s_1 = (s_1^1,...,s_1^{|P|}, v_1, q_1^1,...,q_1^{|C|})$ *and* $s_2 = (s_2^1,...,s_2^{|P|}, v_2, q_2^1,...,q_2^{|C|})$*. We define that* s_2 *is a successor of* s_1*, denoted by* $s_1 \Rightarrow s_2$*, if the following is satisfied:*

- *There exists a process $p_i \in P$ such that (1) for all $j \neq i$ we have that $s_1^j = s_2^j$; and (2)$(s_1^i, e, s_2^i) \in succ$.*
- $(v_2, q_2^1, ..., q_2^{|C|}) = post_e(v_1, q_1^1, ..., q_1^{|C|})$, *where $post_e$ update the valuation of variables from v_1 to v_2 and change message queues contents of respective channels.*

After modelling system, we should describe specification under which the model to be verified. Temporal logic, such as Linear Temporal Logic, which extends proposition logic with temporal operator, is a good choice of description of system properties.

Definition 3 *(Linear Temperoal Logic). Linear Temporal Logic(LTL) has the following syntax given in Backus Naur form:*

$$\phi ::= \bot \mid \top \mid p \mid (p) \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \phi \rightarrow \psi \mid X\phi \mid F\phi \mid G\phi \mid \phi U \psi \mid \phi W \psi \mid \phi R \psi \quad (1)$$

where p is any proposition atom from sone set of atoms. Temporal operator X means next state, F means some future state, and G means all states. The next three, U, W and R, are called Until, Release, *and* Weak-until, *respectively.*

According to the requirement specified in the standard, suppose that there are n BAs altogether, the Mastership Transfer must satisfy the following properties:

1. there cannot be more than one Master at one time; it is written in the LTL as shown in

$$G\neg(BA\_Master_i \wedge BA\_Master_j), i, j = 1, ..., n, i \neq j$$

2. there cannot be no Master at one time; it is written in the LTL as shown in

$$G\neg(\bigwedge_{i=1}^{n} BA\_Standby_i).$$

3 System Modelling

MVBs under Mastership Transfer protocol has two main component to be modelled. In this section, we present our modelling approach for both BusAdmins and the communication mechanism amongst BAs.

3.1 Bus Administrator Modelling

3.1.1 basic data structure

First we define the data structure of Bus Administrator and message frame exchanged amongst different BAs. Bus Administrator data structure defines local variables used for BA itself, including different timers, flag, address, etc. Most important fields of BA structure are *send* and *recv* buffer channel for communicating. These channel contains buffer blocks, each of that consists of both message frame and channel pointer for convenience.

Listing 1. Bus Admin structure

```
typedef BA_DEF
{
   byte T_standby; /*t_standby Timer*/
   byte Turn;  /*turn Timer*/
   bit T_find_next; /*t_find_next Timer*/
   bit T_interim; /*t_interim Timer*/
   bit ACT;    /*BA actualized state flag*/
   byte rank;  /*BA sequence number in BA-
       list*/
   byte adr;   /*BA address in memory*/
   chan send = [10] of {FRAME, chan};
      /*BA sender buffer channel*/
   chan recv = [10] of {FRAME, chan};
      /*BA receiver buffer channel*/
}
```

Listing 2. enumeration of BA state

```
mtype =
{ ba_STANDBY_MASTER,
  ba_REGULAR_MASTER,
  ba_END_OF_TURN,
  ba_FIND_NEXT,
  ba_INTERIM_MASTER}
```

Enumeration of BA state consists of all major distinguished states of BA under Mastership Transfer protocol.

Listing 3. message frame structure

```
typedef FRAME
{
   mtype type; /*Frame type*/
   bit  data; /*Frame data*/
   byte from; /*Source BA ID of frame*/
   byte to; /*Target BA ID of frame*/
}
```

Message frame structure defines type, data, and source and target BA ID of messages frames.

Listing 4. enumeration of message frame type

```
mtype =
{ MASTERSHIP_OFFERED,
  MASTERSHIP_RESPONSE,
  STATUS_REQUEST,
  STATUS_RESPONSE,
  REGULAR}
```

Enumeration of message type consists of all major distinguished states of BA under Mastership Transfer protocol.

3.1.2 finite state machine of Bus Administrator

As defined by CFSM in previous section, our model consists of several concurrency running processes, each of that is an instance of process type BA_FSM, which is identified by argument ID. Each process represent an executing Bus Administrator with a member index by same ID in a BA_DEF structure array.

BA_FSM constructs a finite state machine of Bus Administrator. In each state, BA exercise respective operations, receive or sends message frames and transit to other state on conditions.

Listing 5. BA_FSM process structure

```
proctype BA_FSM(byte ID)
{
    STANDBY_MASTER: /*standby state*/
    ...
    /*acts no master operation: if it
        accepts a Mastership Transfer
        Request or if it detect no Bus
        activity during a time greater than
        t_standby, goto REGULAR_MASTER*/
    REGULAR_MASTER: /*master exercise state
        */
    ...
    /*acts master operation: if it detects
        master conflict, goto STANDBY_MASTER
        ; or if it ends its current turn,
        goto END_OF_TURN*/
    END_OF_TURN: /*end of master turn*/
    ...
    /*looks for next BA in BA-list and send
        Status Request; if BA-list exhausts,
         goto REGULAR_MASTER*/
    FIND_NEXT:  /*find next BA to be master
        */
    ...
    /*gets the Status Response: if the BA is
         configured and actualized, then
        offers mastership to it by sending
        Mastership Transfer Request and goto
         INTERIM_MASTER; else if the BA is
        not actualized or time-out without
        response, goto END_OF_TURN*/
    INTERIM_MASTER: /*interim master state*/
    ...
    /*gets the Mastership Transfer Response
        and goto STANDBY_MASTER. Meanwhile,
        if the response is non-acceptance or
         time-out without response, report
        error*/
}
```

3.2 Communication and Timing Modelling

3.2.1 communication mechanism of BusAdmin

Instead of modelling real Bus component, which may introduce more complex concurrency process, we realize communication between BusAdmins by channels. In initial model setting, each BA has sender and receiver channels both with buffer capacity of 10.

asynchronous communication with buffers By positive capacity buffered channel, BAs communicates with each other asynchronously. These messages receiving and sending actions are similar to pulling and posting mails through intermediary.

- when BA receives(pull) message from *recv* channel, it execute

```
BA[ID].recv ? (temp_frame, BA[ID].
    send)
```

  to get message frame into temporal frame to be parsed; meanwhile, get sender's receive channel pointer as send channel of itself, prepared to be used in following sending actions.

- when BA send(post) message to *send* channel, it execute

```
BA[ID].send ! (temp_frame, BA[ID].
    recv)
```

  to put temporal frame into previous channel pointer(another BA's recv channel); meanwhile, send it's recv channel as channel pointer to be received by the target receiver.

In above communication realization, we make channel as part of communication content, reduce amount of channels used in whole model and alleviate state explorations of states caused by channel numbers.

non-blocking communication When channel is empty or full, respectively, receiving or sending processes block at previous execution statement. Obviously, it is inflexible for modelling more complex system execution situation. So we realized non-blocking communication use *full*, *nfull* and *atomic* primitives.

When sender tries to send, it first confirm whether send channel is full or not, and exclusively satisfies either $full$ or $nfull$ guard condition. If send channel is full, sender process skips sending and execute next statement. If not, sender sends message successfully.

```
do
::full(BA[ID].send)->skip;
::nfull(BA[ID].send)->
   BA[ID].send!(temp_frame,BA[ID].recv);
od;
```

Situation is slightly different for receiver process. We encapsulate receiving execution as atomic segment and provides *else* clause besides. When receiver cannot receive message frame from recv channel as the channel is empty, it will chose *else* bypath and will not be blocked.

```
do
::atomic
   {BA[ID].recv?(temp_frame,BA[ID].send)->
      ...
   }
::else->...
od;
```

3.2.2 timing mechanism

In Mastership Transfer protocol, there are many local timer in each process, such as $T\_standby$, $Turn$, etc. As CFSMs semantics is asynchronous models, different timer in process need to be synchronised.

Many modelling language has synchronous channels, which is different from asynchronous channels, the former can be treated as zero-capacity channel that requires sender and receiver must communicated at the same instant.

Based on the synchronous channel structure, we can make timer in different process has same steps. We model another timing manager process to coordinate different timers. Timer is not updated in each process, but be sent to timing manager by each process before updates through synchronous communication. Then timing manage updates all timer counts in atomic segment and synchronously sends timer back to each process. By timing manager, we realize timers in different process to step by same rate.

4 Property Modelling

After we get system model, we need make specification to be verified. In this section, we first introduce properties classification about our model, then for handling troubles of complex property description and gaining benefits of synthesis, we introduce live sequence chart and combine it with previous system model.

4.1 Property Classification

4.1.1 safety property

A safety property states that some bad thing never happens, representing requirements that should be continuously maintained by the system. To our models, we have following safety property to be verified:

- BusAdmin 0 (BA[0]) and BusAdmin 1 (BA[1]) cannot be masters at same time:

$$G\ mutex \tag{2}$$

where

```
#define mutex
(!(ba0_cur==ba_REGULAR_MASTER &&
    ba1_cur==ba_REGULAR_MASTER))
```

- Once BA[0] becomes master, then system must has a master forever:

$$G(\ ba0\_master \rightarrow G\ !nonemaster)) \tag{3}$$

where

```
#define ba0_master
(ba0_cur==ba_REGULAR_MASTER)
#define nonemaster
(ba0_cur==ba_STANDBY_MASTER &&
 ba1_cur==ba_STANDBY_MASTER &&
 ba2_cur==ba_STANDBY_MASTER)
```

4.1.2 liveness property

A liveness property states that some good thing eventually happens, representing requirements whose eventual realization must be ensured. To our models, we have following liveness property to be verified:

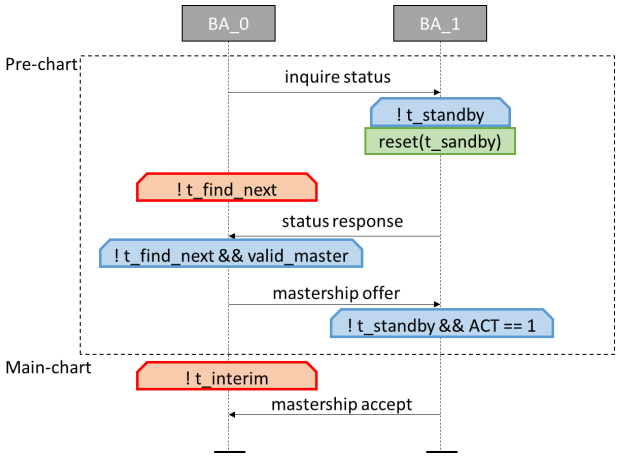


Figure 1. An example of LSC specification: *BA*_0 transfer mas-tership to *BA*_1.

- It is required that BA[0], BA[1] and BA[2] all can infinitely often be the master.

$$GF\ ba0\_master\ \&\&\ GF\ ba1\_master \\ \&\&\ GF\ ba2\_master \tag{4}$$

where $ba1\_master$ and $ba2\_master$ is defined similar as $ba0\_master$.

- After BA[0] becomes the master, it can eventually relieve mastership.

$$G\ (ba0\_master \rightarrow F\ ba0\_standby) \tag{5}$$

where $ba0\_standby$ is defined similar as $ba0\_master$.

4.2 Live sequence charts

Live Sequence Chart (LSC) (Kugler and Segall, 2009; Li et al., 2010) use instance lines, expressions, messages, conditions and updates to describe scenario of interactions among sequence processes. We can iteratively refine spec-ification described by LSCs, that is introduced in initial system design stage and reused for the whole software de-veloping procedures.

Example 1 *In Figure.1, an Live Sequence Chart describes a fragment procedure of mastership transfer from BA_0 to BA_1, which starts from BA_0 send inquire status mes-sage and ends with BA_1 sends mastership accept mes-sage. Red hexagon represents hot condition, which can-not be violated. Blue hexagon represents cool condition, which once be violated, then go back to top and make a new execution.*

Given formal syntax and sematic of LSCs, we can combine normal specification described by LSC with existing formal logics and model checking to verify the properties provided by developer and stakeholder.

Main process of LSC described scenario-based system properties formal description and model checking is that,

1. Use normal LSC to describe process instance interaction scenario in system model, identify the Pre-chart and Main-chart.

2. Translate LSC to Observer Automata *ObsA*. *ObsA* monitors these messages communication, conditions satisfaction and updates consistency proposed by LSC.

3. Combine *ObsA* with original system model $\mathscr{S}$, without introducing side effects in original system operation.

4. Formally, check property

$$G\ (ObsA.l_{min} \rightarrow F\ ObsA.l_{max})$$

where $ObsA.l_{min}$ identifies top of Main-chart, $ObsA.l_{max}$ identities bottom of Main-chart. This property represents that each time the system enters the main chart of LSC, it will eventually reach the bottom of the main chart. Then LSC scenario property is reduced to a classical real-time model checking problem.

4.3 Observer Automata modelling

To translate LSC to observer automata, we need partition LSC into locations, simregions, and cuts firstly. Then a run of observer automata is a successive transitions among states space consisting of cuts and variables valuations.

One requirement for observer automata is that it can monitor concerned information of interactions and configurations in LSC, such as messages. In addition, observer automata need also capture other information such as timers. Meanwhile, observer cannot bring side effects that disturb normal executions of original system model.

We use copy acceptance communication, which pulls a copy message from channel and has no effect on channel information and structure, like

```
BA[ID].recv ? <temp_frame, temp_chan>
```

where syntactic sugar <> means copy message from *recv* to *temp_fram* and *temp_chan*. As mentioned before, timers implement is similar to messages communication, so it is also monitored by observer automata as reference copy.

We define boolean (bit) variables *obsAlmin* and *obsAlmax* as flag denotation for $ObsA.l_{min}$ and $ObsA.l_{max}$ respectively, with initial value of $false(0)$. Based on the above, a monitor process which realizes observer automata *ObsA* and a LSC specification can be both added to our model checking of MVB protocol.

5 Experiments

We implement our approach by explicit states model checking tool SPIN (Holzmann,1997). Its modelling language Promela has all the model elements we proposed previously. To check LSC specification, we implement a tools translate graphic LSC to modelling codes. We run our model checking on 32-bit Window7 platform with 2GB RAM memory limitation. To exhibit the use of our modelling and property setting works

We show the experiments results of MVB protocol model checking against 5 property presented previously in this paper in Table.1.

- First two lines indicates results of first two safety property checking. As safety property is violated by finite prefix of execution, the search depth is equivalent to counterexample depth.

- Lines 3, 4 indicate results of two liveness property checking. Different from above safety property, a counterexample trace that violates liveness property is a infinite suffix of executions, which consists of a lasso structure. Model checking algorithm for search of such lasso is a double-DFS algorithm using stack and has tables. Consequently, these counterexamples depth are less than checking depth reached.

- Last line is result of checking translated LSC property. Obviously, checking a property described by LSC and translated into observer automata is much harder than simple safety and liveness property. It is resulted both from LSC's rich expression and observer automata complex monitoring functions. In this experiment, we can not find counterexample before memory exhausts. Conservatively speaking, the specification described by LSC is satisfied by MVB protocol.

6 Conclusions

Model checking of Multiple Vehicle Bus under mastership transfer protocol shows an industrial verification case study, which combines both device model, communication model and specification model in a unified modelling framework. One benefit of our approach is that, we model Bus administrator as modules, and use channels efficiently. It is an immense improvement on modelling technique compared to (Xia et al., 2013). Another benefit is that we introduce LSCs to make specification of more complex process interaction scenarios. Moreover, with help of translating LSC into observer automata and verifying automata location based liveness property, we can confirm whether LSCs specification is satisfied or not.

We think we have two aspects for future work. One is further improvement of our modelling works and translation from specification to automata. Because the size of models directly affect checking efficiency, we expect smaller models and make possible abstraction. The other extending maybe introduce synthesis technique, which starts from specifications, and seek possible system models that satisfies all the inferred properties.

Table 1. Experiments results of MVB protocol.

Property ID	Transitions	Atomic steps	Memory usage (MB)	Depth reached	Counterexample depth
1	399	127	2.391	966	966
2	347	112	2.391	846	846
3	574767	183018	15.379	1999	1903
4	2964	112	2.586	978	940
5	12201602	3032340	2024.453	1999	—

Acknowledgement. This paper is supported in part by the National Key R&D Program of Chian No. 2017YFC0820100.

References

Edmund M Clarke, Orna Grumberg, and Doron A Peled. *Model checking*. MIT press, 1999.

Gerard J Holzmann. The model checker spin. *Software Engineering, IEEE Transactions on*, 23(5):279–295, 1997.

Jaime Jiménez, Iker Hoyos, Carlos Cuadrado, Jon Andreu, and Aitzol Zuloaga. Simulation of message data in a testbench for the multifunction vehicle bus. In *IEEE Industrial Electronics, IECON 2006-32nd Annual Conference on*, pages 4666–4671. IEEE, 2006.

Hubert Kirrmann and Pierre A Zuber. The iec/ieee train communication network. *Micro, IEEE*, 21(2):81–92, 2001.

Hillel Kugler and Itai Segall. Compositional synthesis of reactive systems from live sequence chart specifications. In *Tools and Algorithms for the Construction and Analysis of Systems*, pages 77–91. Springer, 2009.

Shuhao Li, Sandie Balaguer, Alexandre David, Kim G Larsen, Brian Nielsen, and Saulius Pusinskas. Scenario-based verification of real-time systems using Uppaal. *Formal Methods in System Design*, 37(2-3):200–264, 2010.

Jean-Pierre Queille and Joseph Sifakis. Specification and verification of concurrent systems in cesar. In *International Symposium on Programming*, pages 337–351. Springer, 1982.

Mo Xia, Kueiming Lo, Shuangjia Shao, and Mian Sun. Formal modeling and verification for MVB. *Journal of Applied Mathematics*, 2013, 2013.

Huang Zhiwu, Zhou Sheng, Gui Weihua, and Liu Jianfeng. Research and design of protocol analyzer for multifunction vehicle bus. In *Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on*, pages 8358–8361. IEEE, 2008.

A Model of a Marine Two-Stroke Diesel Engine with EGR for Low Load Simulation

Xavier Llamas Lars Eriksson

Vehicular Systems, Dept. of Electrical Engineering, Linköping University, Sweden
{xavier.llamas.comellas,lars.eriksson}@liu.se

Abstract

A mean value engine model of a two-stroke marine diesel engine with EGR that is capable of simulating during low load operation is developed. In order to be able to perform low load simulations, a compressor model capable of low speed extrapolation is also investigated and parameterized for two different compressors. Moreover, a parameterization procedure to get good parameters for both stationary and dynamic simulations is described and applied. The model is validated for two engine layouts of the same test engine but with different turbocharger units. The simulation results show a good agreement with the different measured signals, including the oxygen content in the scavenging manifold.

Keywords: modeling, parameterization, simulations, exhaust gas recirculation, combustion engines

1 Introduction

The marine shipping industry is facing increased demands in the reduction of harmful exhaust gas emissions. Stricter emission limits of Sulphur Oxides (SOx) and Nitrogen Oxides (NOx) are imposed in certain Emission Control Areas (ECAs). The emission values to fulfill in these ECAs are set by the IMO Tier III limits (International Maritime Organization, 2013) that came into play in January 2016. One of the available technical solutions to achieve the targeted reduction in NOx emissions is Exhaust Gas Recirculation (EGR). An EGR system recirculates a fraction of the exhaust gas into the scavenging manifold, providing burned gases in the combustion chamber that directly decreases the production of NOx during the combustion.

EGR technologies for two-stroke engines are still at the initial phases of its development. In addition, there are not many available vessels with an EGR system installed and thus performing tests is often difficult. Furthermore, testing any new system in marine two-stroke engines is also very costly mainly due to the fuel cost associated with the sizes of such engines. Hence, in order to improve the performance of the EGR control systems, a fast and accurate simulation model is a very valuable tool.

Mean Value Engine Models (MVEMs), are a very common approach for control oriented modeling of internal combustion engines. In particular, EGR systems have been also modeled using this approach. Many interesting research articles about EGR modeling in automotive applications can be found in the literature, some examples are, (Wahlström and Eriksson, 2011; Nieuwstadt et al., 2000). On the other hand, marine two-stroke engines have not been widely studied. Nevertheless, some research papers focused on MVEMs for two-stroke engines are (Blanke and Anderson, 1985; Theotokatos, 2010; Hansen et al., 2013). In addition, in (Guan et al., 2014) the modeling of the low load operation of a two-stroke engine without EGR is studied.

The work presented here is an extension of the model proposed in (Alegret et al., 2015), which enables the model to simulate low engine loads. The low load operation is very relevant for the EGR control since the Tier III emission limits have to be fulfilled near certain coasts, e.g. harbors, where the vessel is normally operating at low loads. The main new component that needs to be introduced for this low load simulation is the auxiliary electrical blower. Its mission is to ensure that there is enough scavenging pressure at low loads when the turbocharger is not capable to provide it. Moreover, the turbocharger model will be required to simulate at low speeds and pressure ratios. This area is normally not measured in the provided performance maps, so a model that can extrapolate to this area is also required.

The developed model is, as in (Alegret et al., 2015), based on the 4T50ME-X test engine from MAN Diesel & Turbo. It is a two-stroke uniflow diesel engine, turbocharged, with variable valve timing and direct injection. Its maximum rated power is $7080\,kW$ at $123\,rpm$. Also, it is equipped with an EGR system and a Cylinder Bypass Valve (CBV).

2 Experimental data

The targeted test engine is constantly being rebuilt to test new components and new control strategies. This implies that it is difficult to find measurement data from the same engine configuration. Most of the measurement data available is from the same layout as the data used in (Alegret et al., 2015). For layout number 1 the oxygen sensors were not properly calibrated and thus cannot be used for validating the oxygen levels at the manifolds. For the model parameterization 30 different stationary points are extracted from the measurement data and another 24 sta-

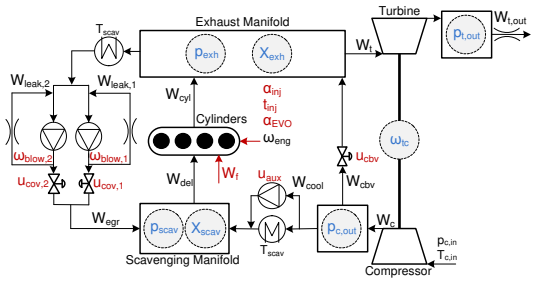


Figure 1. Engine model diagram with states (blue) and control inputs (red).

tionary points are saved for the validation.

Some more data is available from another layout of the engine and will be used for validation of the oxygen level in the scavenging manifold. However, in this layout, number 2, the turbocharger was changed and some sensors where removed. Moreover, there is much less data available, and only 18 stationary points could be extracted for the parameterization and the validation of the model.

3 Modeling

The complete MVEM model consists of thirteen states and nine control inputs. Figure 1 depicts a model diagram. The states are compressor outlet pressure, $p_{c,out}$, scavenging manifold pressure, p_{scav}, exhaust manifold pressure, p_{exh}, turbine outlet pressure, $p_{t,out}$ and turbocharger speed, ω_{tc}. The chemical species mass fractions are states in the scavenging and the exhaust manifolds, X_{scav} and X_{exh}. The species included in the model are oxygen, carbon dioxide, water and sulfur dioxide, $X = [X_{O_2}, X_{CO_2}, X_{H_2O}, X_{SO_2}]$. The dynamic equations are the same for each species so (6) and (7) correspond to eight single ODEs. The dynamic behavior of the modeled states is governed by the following differential equations

$$\frac{d}{dt}\omega_{tc} = \frac{P_t - P_c}{J_t\ \omega_{tc}} \tag{1}$$

$$\frac{d}{dt}p_{c,out} = \frac{R_a\, T_{c,out}}{V_{c,out}}\left(W_c - W_{cool} - W_{cbv}\right) \tag{2}$$

$$\frac{d}{dt}p_{scav} = \frac{R_a T_{scav}}{V_{scav}}\left(W_{cool} + W_{egr} - W_{del}\right) \tag{3}$$

$$\frac{d}{dt}p_{exh} = \frac{R_e\, T_{exh}}{V_{exh}}\left(W_{cyl} - W_{egr} - W_t + W_{cbv}\right) \tag{4}$$

$$\frac{d}{dt}p_{t,out} = \frac{R_e\, T_{t,out}}{V_{t,out}}\left(W_t - W_{t,out}\right) \tag{5}$$

$$\frac{d}{dt}X_{scav} = \frac{R_a T_{scav}}{p_{scav} V_{scav}}\left(X_{exh} - X_{scav}\right) W_{egr} + \frac{R_a T_{scav}}{p_{scav} V_{scav}}\left(X_{amb} - X_{scav}\right) W_{cool} \tag{6}$$

$$\frac{d}{dt}X_{exh} = \frac{R_e T_{exh}}{p_{exh} V_{exh}}\left(X_{cyl} - X_{exh}\right) W_{cyl} \tag{7}$$

The control inputs are EGR blower speeds, $\omega_{blow,1}$ $\omega_{blow,2}$, blower cut-out valves (COV) position, $u_{cov,1}$ $u_{cov,2}$, fuel mass flow, W_f, fuel injection angle α_{inj}, fuel injection time, t_{inj}, exhaust valve closing angle, α_{EVC}, CBV position, u_{cbv}, and auxiliary blower operation u_{aux}. Engine speed, ω_{eng}, and compressor inlet pressure and temperature, $p_{c,in}$ $T_{c,in}$, are considered known inputs to the model.

It is simple to reduce the model to seven states if we are only interested in tracking the oxygen level in the manifolds. Then in (6) and (7), the mass fraction only refers to oxygen, e.g. $X = X_{O_2}$. The model is built using different submodels interconnected. The submodels are mainly control volumes, e.g. scavenging manifold, and flow elements e.g. cylinder bypass valve, compressor, turbine, etc. Since the model is an extension of the one described in (Alegret et al., 2015), only the new or modified submodels are presented here.

3.1 Compressor

In order to properly simulate low loads, a compressor model capable of predicting mass flow and efficiency at low speeds is required. The chosen compressor mass flow model is the one developed in (Leufvén and Eriksson, 2014), which is capable of this extrapolation. In addition, the proposed model is capable of predicting mass flows down to pressure ratios below one and to zero compressor speed. The area where the compressor normally operates during low load is below the slowest measured speed line, which is depicted in Figures 2 and 3.

In the model, each compressor speed line is described by a super ellipse, which mathematically is written as

$$\left(\frac{\bar{W}_c - \bar{W}_{ZS}}{\bar{W}_{Ch} - \bar{W}_{ZS}}\right)^{CUR} + \left(\frac{\Pi_c - \Pi_{Ch}}{\Pi_{ZS} - \Pi_{Ch}}\right)^{CUR} = 1 \tag{8}$$

where $\bar{W}_{ZS}$, $\bar{W}_{Ch}$, Π_{ZS}, Π_{Ch} and CUR are functions of compressor speed. More details about these functions and the model in general can be found in (Leufvén and Eriksson, 2014). Since (8) is invertible, it can be used to predict either pressure ratio given compressor speed and mass flow or mass flow given compressor speed and pressure ratio. The latter case is used in the proposed engine model.

The compressor efficiency is modeled using the ideas from (Martin et al., 2009) for the isentropic efficiency definition. The key for the model is to use the Euler's equation (Dixon and Hall, 2013) applied to the compressor velocity triangles. Using the simplifications from (Martin et al., 2009), the conclusion is that the actual enthalpy rise for a fixed compressor speed can be modeled as a linear function of the mass flow. This simplifies the number of parameters required and since it is based in the physical equations, makes the extrapolation to the low load area more reliable. The proposed compressor model requires 15 parameters for the mass flow submodel and 4 parameters for the efficiency submodel.

The mass flow model together with the efficiency model fitted to the two compressors used in this study is depicted in Figures 2 and 3. The absolute value of the relative errors for both compressors are shown in Table 1.

3.2 Turbine

The Turbine mass flow model is very similar to the one used in (Alegret et al., 2015). However, a modification in the model is required to describe the turbine speed depen-

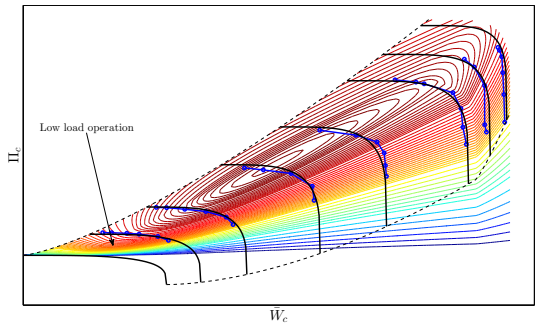


Figure 2. Compressor 1 model, in black, plotted together with the measured map points, in blue dots. The first speed line in the lower left corner represents the stand still characteristics of the compressor. The thinner level lines represent the modeled efficiency extrapolation down to unity pressure ratio.

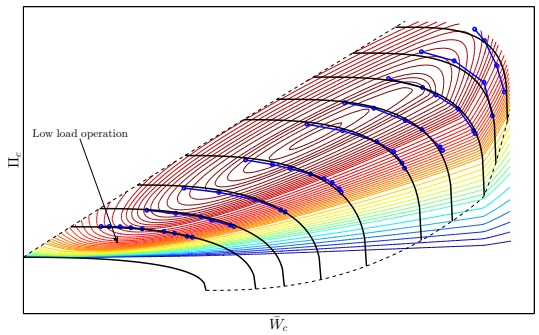


Figure 3. Compressor 2 model, in black, plotted together with the measured map points, in blue dots. The first speed line in the lower left corner represents the stand still characteristics of the compressor. The thinner level lines represent the modeled efficiency extrapolation down to unity pressure ratio.

dence observed for Turbine 2. The mass flow is described by the following function from (Eriksson and Nielsen, 2014)

$$\bar{W}_t = C_t \sqrt{1-(\Pi_t+\Pi_0)^{k_t}} \tag{9}$$

where k_t, Π_0 and C_t are constant parameters to be estimated for Turbine 1. For Turbine 2, k_t and C_t are also constants but Π_0 is modeled using a quadratic polynomial of corrected turbine speed to capture the different speed lines observed in Figure 5. The quadratic polynomial is defined as

$$\Pi_0 = c_{\Pi_0,1}\bar{N}_t^2 + c_{\Pi_0,2}\bar{N}_t + c_{\Pi_0,3} \tag{10}$$

where $c_{\Pi_0,1}$, $c_{\Pi_0,2}$ and $c_{\Pi_0,3}$ are model parameters. The two models fitted to the two turbines are shown on the left side of Figures 4 and 5.

Table 1. Absolute value of the relative errors (%) for both compressors and turbines. Mean indicates the mean value of all errors while Max. is the maximum error computed.

	Compressor 1		**Compressor 2**		**Turbine 1**		**Turbine 2**	
	Mean	Max.	Mean	Max.	Mean	Max.	Mean	Max.
$\bar{W}$	0.86	2.21	0.59	1.87	0.14	0.51	0.56	5.61
η	1.47	4.51	0.74	2.87	0.60	3.32	0.69	6.11

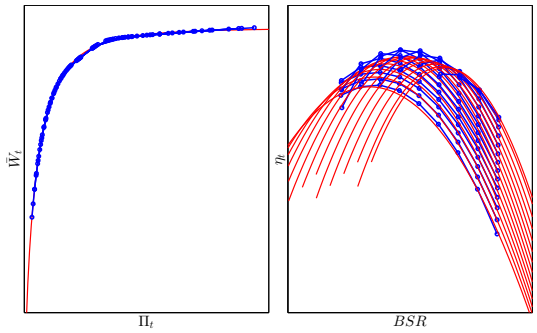


Figure 4. Left: turbine 1 mass flow model, in red, plotted together with the measured map points, in blue dots. Right: turbine 1 efficiency model, in red, plotted with the measured efficiency points, in blue dots.

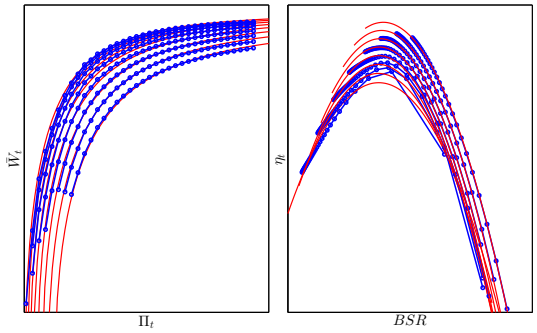


Figure 5. Left: turbine 2 mass flow model, in red, plotted together with the measured map points, in blue dots. Right: turbine 2 efficiency model, in red, plotted with the measured efficiency points, in blue dots.

The turbine efficiency is modeled using the Blade Speed Ratio (BSR), as in (Wahlström and Eriksson, 2011; Alegret et al., 2015). The relation between turbine efficiency and BSR is defined as

$$\eta_t = c_1 BSR^2 + c_2 BSR + c_3 \tag{11}$$

where c_1, c_2 and c_3 are also quadratic functions of turbine corrected speed, and each polynomial is defined as follows

$$c_X = c_{X,1}\bar{N}_t^2 + c_{X,2}\bar{N}_t + c_{X,3} \tag{12}$$

In total the turbine efficiency consists of nine parameters. The modeled and the measured efficiencies for both turbines are depicted in Figures 4 and 5. Furthermore, for both turbines, the absolute value of the relative errors are presented in Table 1.

3.3 EGR Blowers

The EGR blowers are modeled as in (Alegret et al., 2015). The only difference is that there are two equal blowers in parallel. The EGR flow is controlled with the blower speed control inputs, $\omega_{blow,1}$ and $\omega_{blow,2}$, and the cut-out valves are used to open or close the EGR flow. Depending on the engine running mode, there can be only one blower operating or both of them if more EGR flow is required.

3.4 Auxiliary Blower

The pressure increase of an electric blower is often modeled as a quadratic function of the volumetric flow (Guan et al., 2014). Since the blower's pressure increase is available from the system states, the quadratic function is inverted to obtain the volumetric flow. Using the pressure and temperature at the inlet the mass flow provided by the blower is obtained

$$W_{Aux} = \frac{p_{c,out}\left(c_{Aux,1} + c_{Aux,2}\sqrt{c_{Aux,2} - (p_{scav} - p_{c,out})}\right)}{R_a T_{c,out}} \quad (13)$$

where $c_{Aux,1}$, $c_{Aux,2}$ and $c_{Aux,3}$ are tuning parameters estimated with the blower technical specifications.

For the studied test engine, the auxiliary blower is installed in parallel with the air cooler after the compressor. This means that when operating it will pull air mass flow from the compressor outlet control volume to the scavenging manifold. When active, the pressure difference between these two control volumes will reverse. When the auxiliary blower is operated (since there is no restriction valve for reverse flow in the cooler), there is flow recirculation in the air cooler. This issue needs to be modeled in order to capture the measured system behavior. Thus, the flow from the compressor outlet to the scavenging manifold is then modeled using (13) and two incompressible flow restrictions from (Eriksson and Nielsen, 2014)

$$W_{cool} = \begin{cases} W_{Aux} - A_{cool,r}\sqrt{\frac{p_{scav}\left(p_{scav} - p_{c,out}\right)}{T_{scav}}} & \text{if } u_{aux} = 1 \\ A_{cool}\sqrt{\frac{p_{c,out}\left(p_{c,out} - p_{scav}\right)}{T_{c,out}}} & \text{if } u_{aux} = 0 \end{cases} \quad (14)$$

where A_{cool} represents the flow restriction when the blower is inactive and $A_{cool,r}$ models the magnitude of the recirculation when the blower is running. Both are parameters to be estimated.

3.5 Exhaust Back Pressure

An incompressible flow restriction together with a control volume is used to model the back pressure for the turbine, $p_{t,out}$. The pressure dynamics are described by (5). And the exhaust flow $W_{t,out}$ is modeled using the standard incompressible flow restriction from (Eriksson and Nielsen, 2014) where the restriction area is a tuning parameter.

3.6 Combustion Species and Thermodynamic Parameters

The species mass fraction out of the cylinders, X_{cyl}, is calculated using the stoichiometric combustion equation and the air and fuel flows entering the cylinders. Without including the nitrogen explicitly, the combustion equation can be written

$$CH_yS_z + (1 + y/4 + z)O_2 = CO_2 + (y/2)H_2O + zSO_2 \quad (15)$$

where y is the hydrogen to carbon ratio and z the amount of sulfur in the fuel, which are known parameters. In the case of the reduced model with only oxygen mass fraction, X_{cyl} can be computed as in (Alegret et al., 2015; Wahlström and Eriksson, 2011).

Furthermore, the species vector is used to compute the thermodynamic parameters, R, c_p and γ of the working gas. This is done for each different gas composition and computed together with the gas temperature using the Nasa polynomials that can be found in (Goodwin et al., 2014).

4 Parameterization Procedure

The parameterization is done in similar steps as it is described in (Alegret et al., 2015). First the following submodels are parameterized alone: compressor, turbine, ERG blowers and Aux blower. These submodel parameters are kept fixed in the following parameterization steps.

4.1 Complete stationary parameterization

The path to follow would be to estimate the different flow restrictions of the model independently and then do a complete parameterization of the whole model together. However, this is not possible since there is no mass flow measurement available. Hence, the next step in the parameterization is to use the complete model to get the best set of parameters that predict the measured states.

The method followed here differs from the one used previously in (Alegret et al., 2015) where the derivatives of the states are used in the parameterization. Here instead the whole model is simulated at each stationary point and the simulated stationary states are used to compute the relative errors, e_{rel}. Finally, a least-squares problem is formulated with the following objective function

$$V_{stat}(\theta) = \frac{1}{NS}\sum_{i=1}^{S}\sum_{n=1}^{N}(e^i_{rel}[n])^2 \quad (16)$$

where S is the number of different measured signals used to compute the relative errors, in the general case those signals are: $[p_{scav}, p_{exh}, p_{c,out}, p_{t,out}, \omega_{tc}, T_{exh}, P_{eng}, W_{egr}]$. N is the number of stationary points used. The vector θ represents the parameters to be estimated, which in this case are all the static parameters, except of the fixed parameters of the submodels stated in the beginning of Section 4. The stationary simulations are done using a Matlab/Simulink implementation of the model. To reduce the computational time required, the Matlab parallel computing toolbox is used to run simultaneous simulations. Furthermore, a check on the state derivatives is done in order to stop the simulation once the stationary levels are reached.

Different estimation steps are done to ensure that the solver does not get lost with too many parameters. The parameterization is started without EGR, CBV or low load stationary points which are progressively included in the successive steps. Also, the results from each parameterization step are used as initial guess for the following one. Finally, a complete parameterization with all stationary points available is done.

Table 2. Absolute value of the model relative errors (%) for both engine layouts. Low load is with Aux. blower active, Mid load is below 70 % and High load is above 70 %

	Engine Layout 1							
	p_{scav}	p_{exh}	$p_{c,out}$	$p_{t,out}$	ω_{tc}	T_{exh}	$P_{eng,i}$	W_{egr}
Low Load	2.39	2.62	5.13	2.07	2.80	0.80	0.44	11.44
Mid Load	3.51	3.36	6.86	3.37	1.90	0.78	0.36	9.84
High Load	4.74	6.92	2.05	5.29	2.58	1.28	2.52	5.55
	Engine Layout 2							
	p_{scav}	p_{exh}	$p_{c,out}$	$p_{t,out}$	ω_{tc}	T_{exh}	X_{O_2}	-
Low Load	4.75	7.73	4.38	0.17	6.32	1.74	1.93	-
Mid Load	3.75	3.08	4.78	0.18	2.26	1.59	1.03	-
High Load	7.69	10.48	4.42	0.48	2.35	3.48	0.89	-

4.2 Dynamic estimation

Fixing the estimated parameters at the previous steps, the dynamic parameters, J_t, V_{scav}, V_{exh}, $V_{c,out}$, $V_{t,out}$ and τ_{cov}, are tuned using the same procedure as in (Alegret et al., 2015). In this case, 17 different step responses are used, including tree load steps with the auxiliary blowers active.

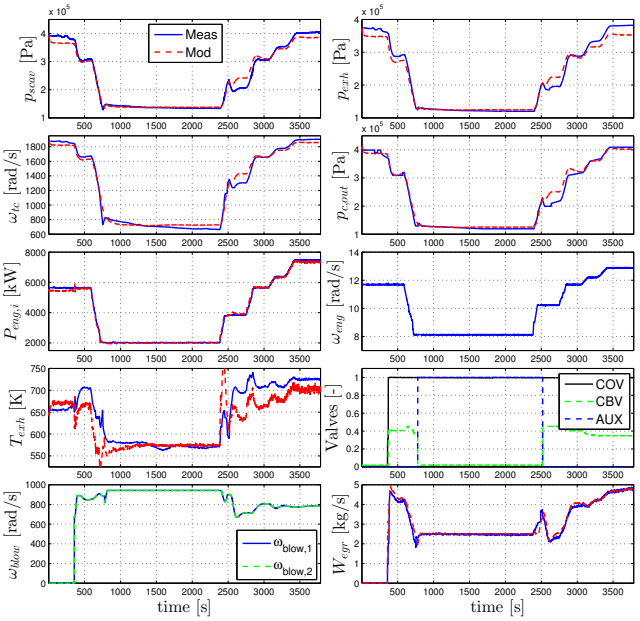


Figure 6. Model simulation vs measurements for Engine Layout 1.

5 Model Validation

Table 2 presents the absolute value of the relative errors separated for different load ranges. For Layout 1, the 24 validation stationary points are used. For Layout 2, since there is few data available, the errors are computed with the same stationary points used in the estimation. For both Layouts, the higher pressure errors are mostly in the high load case, where also the exhaust temperature and indicated engine power errors are higher. This indicates that the model and in particular the Seiliger cycle could be improved in this area. One reason for this could be that at high load is where the engine protection controls limit the maximum pressure in the cylinders, and this might not be totally captured in the model. On the other hand, for the mid and low load ranges the errors are in general of similar magnitude. Note that for the Layout 2, the engine power and the EGR mass flow measurements are not available.

Figure 6 shows the simulation results of Layout 1 compared to the measurements of a dataset not used in the estimation. This simulation has a low load phase where the auxiliary blower is enabled, where it can be seen that the system behavior is captured by the model. In particular the measured pressure and turbocharger speed values are matched by the model. More discrepancy is observed in the modeled exhaust temperature during the transients.

For Layout 2, the simulation results are presented in Figure 7. There is also a low load operation that the model is capable to capture. In this case the turbocharger speed prediction is worse than for the Layout 1 which in turn affects the stationary levels of the pressures. Nevertheless, it is important to mention that it has been parameterized with few data. Oxygen mass fraction validation could not be done due to unreliable measurements for Layout 1 and in the previous investigations from (Alegret et al., 2015). Therefore, the most relevant result from Layout 2 is that the scavenging oxygen level is captured as it can be seen in Table 2 and in Figure 7.

6 Conclusions

An MVEM for a marine two-stroke diesel engine capable of simulating low loads is proposed and validated. A parameterization method to overcome the lack of mass flow measurements is also proposed. The main characteristic is that it uses a Simulink model to integrate the modeled states with stationary inputs and is used to compute the residuals. Two different layouts of the same engine but with different turbochargers are investigated, the results show a good agreement between simulation results and measurements with some room for improvement at high loads. The oxygen prediction capabilities are also validated for the second engine layout since the oxygen measurements are reliable.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 634135.

MAN Diesel & Turbo and in particular Guillem Alegret are also acknowledged for their support and suggestions about the modeled test engine.

A Nomenclature

References

Guillem Alegret, Xavier Llamas, Morten Vejlgaard-Laursen, and Lars Eriksson. Modeling of a large marine two-stroke diesel engine with cylinder bypass valve and EGR system.

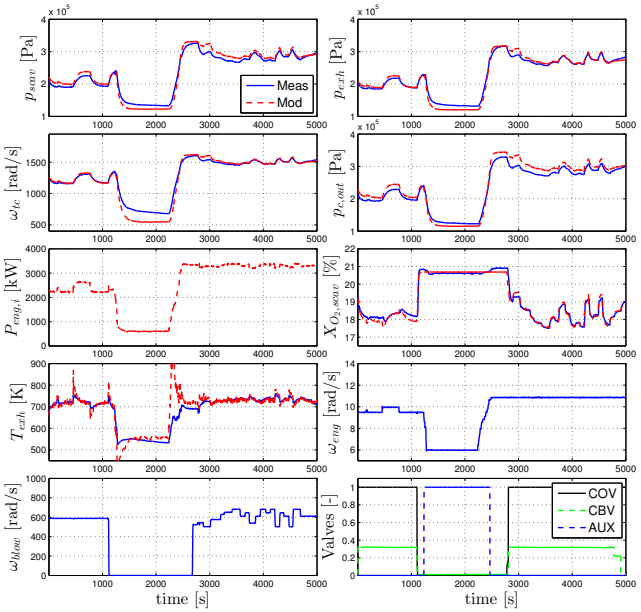


Figure 7. Model simulation vs measurements of for Engine Layout 2.

Table 4. List of symbols

A	Area	$[m^2]$
c_p	Specific heat at constant pressure	$[J/(kgK)]$
J	Inertia	$[kg\,m^2]$
W	Mass flow	$[kg/s]$
$\bar{W}$	Corrected mass flow	$[kg/s]$
p	Pressure	$[Pa]$
P	Power	$[kW]$
R	Gas constant	$[J/(kgK)]$
T	Temperature	$[K]$
V	Volume	$[m^3]$
X	Mass fraction	$[-]$
α	angle	$[rad]$
γ	Specific heat capacity ratio	$[-]$
η	Efficiency	$[-]$
Π	Pressure ratio	$[-]$
$\bar{N}$	Corrected rotational speed	$[rpm]$
ω	Rotational speed	$[rad/s]$

Table 3. Subscripts

cov	cut-out valve	*inj*	injection
blow	blower	*meas*	measured
c	compressor	*mod*	modeled
scav	scavenging manifold	*cool*	cooler
cyl	cylinder	*t*	turbine
del	delivered	*eng*	engine
e	exhaust gas	*a*	air
tc	turbocharger	*x,in*	inlet of x
aux	auxiliary blower	*x,out*	outlet of x
exh	exhaust manifold	*egr*	EGR gas

IFAC-PapersOnLine, 48(16):273 – 278, 2015. ISSN 2405-8963. doi:http://dx.doi.org/10.1016/j.ifacol.2015.10.292. 10th IFAC Conference on Manoeuvring and Control of Marine Craft MCMC.

M. Blanke and J. A. Anderson. On modelling large two stroke diesel engines: new results from identification. *IFAC Proceedings Series*, pages 2015–2020, 1985.

Sydney Lawrence Dixon and Cesare A Hall. *Fluid Mechanics and Thermodynamics of Turbomachinery*. Butterworth-Heinemann, 7th edition, 2013.

Lars Eriksson and Lars Nielsen. *Modeling and Control of Engines and Drivelines*. John Wiley & Sons, 2014.

David G. Goodwin, Harry K. Moffat, and Raymond L. Speth. Cantera: An object-oriented software toolkit for chemical kinetics, thermodynamics, and transport processes. `http://www.cantera.org`, 2014. Version 2.1.2.

C. Guan, G. Theotokatos, P. Zhou, and H. Chen. Computational investigation of a large containership propulsion engine operation at slow steaming conditions. *Applied Energy*, 130: 370–383, 2014. doi:10.1016/j.apenergy.2014.05.063.

Jakob Mahler Hansen, Claes-Göran Zander, Nikolai Pedersen, Mogens Blanke, and Morten Vejlgaard-Laursen. Modelling for control of exhaust gas recirculation on large diesel engines. *Proceedings of the 9th IFAC conference on Control Applications in Marine Systems*, 2013.

International Maritime Organization. *MARPOL: Annex VI and NTC 2008 with Guidelines for Implementation*. IMO, 2013. ISBN 978-92-801-15604.

Oskar Leufvén and Lars Eriksson. Measurement, analysis and modeling of centrifugal compressor flow for low pressure ratios. *Int. J. Engine Res.*, pages 1–16, December 2014. doi:10.1177/1468087414562456.

Guillaume Martin, Vincent Talon, Pascal Higelin, Alain Charlet, and Christian Caillol. Implementing turbomachinery physics into data map-based turbocharger models. *SAE Int. J. of Engines*, 2(1):211–229, April 2009. doi:10.4271/2009-01-0310.

M.J. Nieuwstadt, I.V. Kolmanovsky, P.E. Moraal, A. Stefanopoulou, and M. Jankovic. EGR-VGT control schemes: experimental comparison for a high-speed diesel engine. *IEEE Control Systems Mag.*, 20(3):63–79, 2000.

G. Theotokatos. On the cycle mean value modelling of a large two-stroke marine diesel engine. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of engineering for the maritime environment*, 224(3):193–206, 2010. ISSN 1475-0902.

J. Wahlström and L. Eriksson. Modelling diesel engines with a variable-geometry turbocharger and exhaust gas recirculation by optimization of model parameters for capturing non-linear system dynamics. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 225:960–986, 2011.

Safe Active Learning of a High Pressure Fuel Supply System

Mark Schillinger[1] Benedikt Ortelt[1] Benjamin Hartmann[1]
Jens Schreiter[2] Mona Meister[2] Duy Nguyen-Tuong[2]
Oliver Nelles[3]

[1]Bosch Engineering GmbH, Germany, mark.schillinger@de.bosch.com
[2]Robert Bosch GmbH, Germany
[3]Automatic Control, Mechatronics, Department of Mechanical Engineering, University of Siegen, Germany

Abstract

When modeling technical systems as black-box models, it is crucial to obtain as much and as informative measurement data as possible in the shortest time while employing safety constraints. Methods for an optimized online generation of measurement data are discussed in the field of Active Learning. Safe Active Learning combines the optimization of the query strategy regarding model quality with an exploration scheme in order to maintain user-defined safety constraints. In this paper, the authors apply an approach for Safe Active Learning based on Gaussian process models (GP models) to the high pressure fuel supply system of a gasoline engine. For this purpose, several enhancements of the algorithm are necessary. An online optimization of the GP models' hyperparameters is implemented, where special measures are taken to avoid a safety-relevant overestimation. A proper risk function is chosen and the trajectory to the sample points is taken into account regarding the estimation of the samples feasibility. The algorithm is evaluated in simulation and at a test vehicle.

Keywords: machine learning, system identification, active learning, Gaussian process models, automotive applications

1 Introduction

For calibration purposes, models are used in order to speed up the calibration process, reduce the risk of damages of the system and reduce the time the real system needs to be available. For example, these models can be used in Hardware-in-the-Loop or Software-in-the-Loop environments, or even for automatic model based controller tuning. These models can either be constructed exploiting physical principles (white-box modeling), using data-based modeling techniques (black-box modeling) or a combination of the former (gray-box modeling). White-box models are often hard to generate, as the physical principles and parameters are frequently very complex, hard to model or unknown. Black-box and gray-box modeling overcomes these disadvantages, but strongly depends on informative measurement data. One approach for gathering this measurement data while keeping safety constraints and estimating a black-box model is presented in this paper. We apply the approach to the high pressure fuel supply system (HPFS system) of a gasoline engine.

Active Learning is a subfield of machine learning. Its main idea is to employ a learning algorithm, which chooses queries to be labeled on its own. Labeling data for model learning is often costly, for example due to necessary manpower or expensive test bench time. Thus, it is beneficial to optimize the queries to be labeled in order to minimize the amount of necessary data to achieve a sufficient model quality. In (Settles, 2009) an overview about the Active Learning topic and the corresponding literature is provided.

When taking measurements of a technical system, it is essential to guaranty the integrity of the system. Especially in the case of open-loop measurements, critical system states can occur when choosing improper input signals. For example, when measuring the HPFS system, the maximum allowed rail pressure may not be exceeded.

In many cases, an automation system will avoid threshold violations. Nonetheless, even the attempt to measure unsafe input signals yields side effects: high stress on the test subject, wasted measurement time, and the risk of emergency shutdowns during automated measurement runs. Thus, it is beneficial to avoid critical system inputs in advance.

One possibility to combine the goals of Active Learning, i. e. learning the best model possible using the smallest amount of data, with the goal of avoiding unsafe input queries is presented in (Schreiter et al., 2015). There, a Safe Active Learning algorithm (SAL algorithm) based on GP models is introduced, which finds a set of input samples maximizing the model's entropy, provided that an estimated safety measure is satisfied.

In this paper, the algorithm proposed in (Schreiter et al., 2015) is translated to the real-world application of the HPFS system of a gasoline engine. Compared to the theoretical investigations in (Schreiter et al., 2015), the algorithm is modified in order to meet the requirements of our system. To the best of the authors' knowledge, this is the first time the algorithm is evaluated at a real-world system.

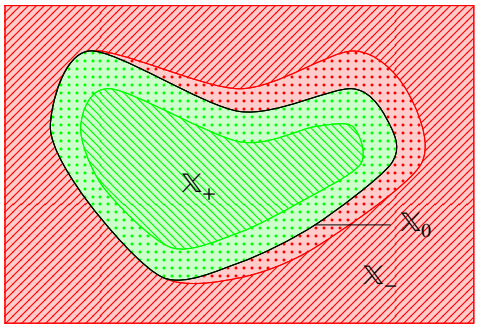


Figure 1. Partition of the input space $\mathbb{X}$ into a safe explorable area $\mathbb{X}_+$ and an unsafe region $\mathbb{X}_-$ separated by the unknown decision boundary $\mathbb{X}_0$. The figure is taken from (Schreiter et al., 2015). There, a discriminative function is learned over the dotted area for recognizing whether the exploration becomes risky.

Table 1. Overview of the models used in SAL

	regression model	classifier
model output	f	g
training data	$y(\boldsymbol{x}_i)$	$h(\boldsymbol{x}_i)$ or $c(\boldsymbol{x}_i)$
SAL's goal	max. model's entropy	satisfy safety constraints

This contribution is organized as follows: First, the fundamentals of the SAL algorithm and the HPFS system are presented. In Section 3, we describe our enhancements of the algorithm and the necessary design decisions. Subsequently, the evaluation in simulation and at a test vehicle is shown.

2 Fundamentals

In this section, we will commemorate the fundamentals of SAL as introduced in (Schreiter et al., 2015) and describe the HPFS system considered in this paper.

2.1 Safe Active Learning

The key goals of SAL according to (Schreiter et al., 2015) are:

1. approximate the system based on sampled data as informative as possible,
2. use a limited budget of measured points, and
3. ensure that critical regions of the considered system are avoided during the measurement process.

A compact and connected input space $\mathbb{X} \subset \mathbb{R}^d$ is defined that is divided into two subspaces $\mathbb{X}_+$ and $\mathbb{X}_-$ in which the system is safe or unsafe, respectively (compare Figure 1). The latter should be avoided with a probability higher than the user defined threshold δ.

In this input space, two GP models are learned. One of the models is a regression of the system output y. For this model, a usual GP is used, which is trained using noisy observations $y = f + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The second model is a problem specific classifier, used to estimate the boundary $\mathbb{X}_0$. It learns a discriminative function $g : \mathbb{X} \to \mathbb{R}$, mapped to the unit interval to describe the class likelihood for each point. This model is trained with class labels $c(\boldsymbol{x}_i) \in \{-1, +1\}$ or discriminative function values $h(\boldsymbol{x}_i) \in (-1, 1)$, depending on the location of the measured point $\boldsymbol{x}_i$ in $\mathbb{X}$ (compare Figure 1). Thereby it is possible either to only use information whether a point was feasible or not, or to use more detailed data about the grade of its feasibility. Here, the discriminative function value is given by $h = g + \zeta$, where $\zeta \sim \mathcal{N}(0, \tau^2)$ specifies the noise. The mixed kind of training data results in a non-Gaussian model likelihood, hence a Laplace approximation is required, to calculate the model's posterior.

Both GP models use zero mean centered Gaussian priors and squared exponential covariance functions. A major prerequisite is that the hyperparameters of both models need to be known in advance. These are the signal magnitude $\sigma_\bullet^2$, the length-scales $\boldsymbol{\lambda}_\bullet \in \mathbb{R}^d$, and the noise variance σ^2 or τ^2. They are summarized in $\boldsymbol{\theta}_\bullet$. The dot $\bullet$ is a placeholder for the regression and discriminative model, specified by index f or g subsequently. An overview of the two models is given in Table 1.

In SAL, the next measurement point $\boldsymbol{x}_{i+1}$ is selected based on a differential entropy criterion. Thus, $\boldsymbol{x}_{i+1}$ is chosen such that the entropy of the regression model f increases as much as possible. Namely, the variance at $\boldsymbol{x}_{i+1}$ is maximized. In order to ensure safety, the optimization is constrained with a safety criterion depending on the predicted probability of failure from the discriminative model. For more details, we refer to (Schreiter et al., 2015).

2.2 The high pressure fuel supply system

The main components of a HPFS system are the high pressure rail, the high pressure fuel pump, and the ECU (Engine Control Unit; compare Figure 2). The pump is actuated by the crankshaft of the engine. A demand control valve in the pump allows to control the delivered volume per stroke. A pressure-relief valve is also included in the pump, but should never open if possible. Hence, we want to limit the maximum pressure during the whole measurement process. The pump transports the fuel to the rail, which contains the pressure sensor. From there, it is injected into the combustion chambers. See (Robert Bosch GmbH, 2005) for more details.

The system has three inputs and one output. The engine speed ($nmot$) affects the number of strokes per minute of the pump and the engine's fuel consumption. The fuel pump actuation (MSV) gives the fuel volume which is transported with every stroke of the pump. It is applied by opening and closing the demand control valve accordingly during one stroke of the pump. The injection time is a variable calculated by the ECU, which sums up the opening times of the single injectors and is, thus, related to the discharge of fuel from the rail.

A notable difference to the systems considered in

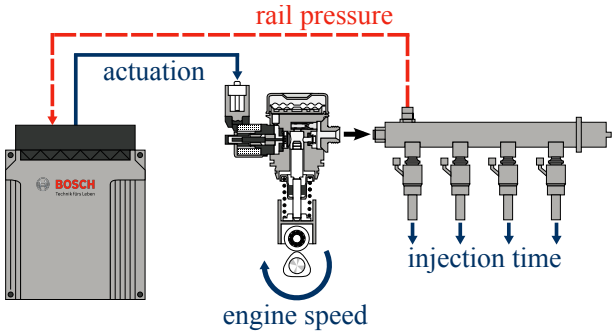


Figure 2. Sketch of the HPFS system's main components, inputs (continuous lines), and output (dashed line). The figure is taken from (Tietze et al., 2014).

(Schreiter et al., 2015) is that the input signals cannot be set directly, but have to be changed continuously. This becomes clear using the example of the engine speed, which cannot change immediately. Another reason for changing the input signals slowly is the system's dynamic behavior. Too fast actions could lead to pressure overshoots which can damage the engine.

If we move from each measurement point on the shortest path to the next one, we would have to assume that $\mathbb{X}_+$ is convex. However, experiments show that this assumption does not hold in the case of the HPFS system. Thus, we always make the detour via a global safe point (GSP), which weakens the precondition on $\mathbb{X}_+$ to star-convexity. This is a sufficient approximation of its real shape if the GSP is chosen correctly. When predicting the feasibility of an input point, the trajectory from the GSP to that point needs to be considered as well.

During the measurement procedure, the injection time is not varied manually but set by the ECU. The permissible times depend on many factors and a wrong choice could extinguish the combustion or even damage components. The engine load would have a major influence on the HPFS system via the injection time, but is omitted to prevent the necessity of a vehicle test bench. In principle, both variables could be additionally considered in the SAL algorithm without major changes of the method.

3 Design and Implementation

The system considered in this paper is defined as follows: the d-dimensional input $\boldsymbol{x} \in \mathbb{X}$ results in a one-dimensional output $y \in \mathbb{R}$ to be modeled. As the input space is divided in a safe and unsafe subspace (compare Section 2.1), we are only interested in measuring the system output $y : \mathbb{X}_+ \to \mathbb{R}$. The safe and unsafe subspace must be distinguishable by supervising the d_z-dimensional additional output $\boldsymbol{z} \in \mathbb{R}^{d_z}$. We assume that all outputs are only observable within $\mathbb{X}_+$. Outside of this subspace, the system cannot be operated safely and thus, no steady state measurements can be taken.

In order to apply the SAL algorithm presented in (Schreiter et al., 2015) to the HPFS system of a real car, three main issues have to be solved:

- learning the hyperparameters $\boldsymbol{\theta}_f$ of the regression model and $\boldsymbol{\theta}_g$ of the discriminative model,
- defining the risk function $\tilde{h} : \boldsymbol{z} \to [-1, 1]$ based on the supervised system output, which is, in combination with measured data, used to calculate the discriminative function value $h = \tilde{h}(\boldsymbol{z})$, and
- implementing the assessment of the feasibility of the trajectory to the next sample.

In the progress of solving these issues, several changes of the algorithm become necessary. This includes:

- implementing an online estimation of the hyperparameters, which requires carefully chosen limits,
- finding a heuristic for an initial set of hyperparameters, and
- replacing the problem specific classifier applied in (Schreiter et al., 2015), which can be trained with labels as well as discriminative function values, by a usual GP regression model.

In the following, these issues are discussed.

3.1 Training of the hyperparameters

In (Schreiter et al., 2015) it is assumed that the hyperparameters of the GP models are given in advance. Thus, when modeling a system from scratch, the hyperparameters need to be determined before the SAL algorithm can be started. The hyperparameters are usually learned by maximizing the marginal likelihood, as shown in (Rasmussen and Williams, 2006). Therefore, already observed data is required. If we assume that we only know one starting point in $\mathbb{X}_+$ in advance, we have almost no idea where the system is in safe operation, so that it is not possible to safely generate measurement data to estimate the hyperparameters.

If we have expert knowledge, which gives us a sufficiently large subspace of $\mathbb{X}_+$, we could generate and measure a space-filling design of experiment (DoE) in this subspace. With the generated data, we could estimate the hyperparameters and subsequently start the SAL algorithm. In Section 4.1, we benchmark this approach against the online hyperparameter training we will develop in the following. A representative subspace of $\mathbb{X}_+$ and a suitably chosen number of predefined measurement points are necessary to obtain a good model with little measurement data. This contrasts the goals of SAL to model the system with little initial knowledge about $\mathbb{X}_+$ and to be content with little measurement data. In industrial practice, a system of similar complexity as the HPFS system would be modeled using about 25 measurement points. Thus, the necessity of previous hyperparameter estimation is a major drawback of the original SAL algorithm.

To overcome this drawback, we estimate the hyperparameters during the SAL. We have to be very careful doing so, as falsely estimated hyperparameters can lead to an overestimation of $\mathbb{X}_+$ and hence to samples in $\mathbb{X}_-$. The estimation is especially error-prone if only a small number of samples is available yet. To reduce this risk, we limit the classifiers length-scales λ_g during the first optimization steps.

According to (Rasmussen and Williams, 2006), the characteristic length-scales λ of a Gaussian process with squared exponential kernel can be interpreted as the distances one has to move in the input space, before the function can change significantly. Thus, large length-scales λ_g of the classifier will result in a fast exploration, because the SAL algorithm assumes little changes in the discriminative function. We do not want the length-scales of the classifier to become too large, as this gives raise to overestimate $\mathbb{X}_+$. To prevent this, λ_g is limited to $\frac{\Delta x}{4}$ until 10 points have been measured and to $\frac{\Delta x}{2}$ until 20 measurement samples have been acquired, where $\boldsymbol{\Delta x} \in \mathbb{R}^d$ is the extent of $\mathbb{X}$ in each input dimension. During the first 5 steps, λ_g is kept constant at $\frac{\Delta x}{4}$, to ensure a good conditioning of the hyperparameter optimization problem. With increasing number of available samples, the estimation of the hyperparameters improves and the limits can be relaxed.

This setup was chosen heuristically and shows good results in the practical application. A motivation for this choice, though no formal derivation, can be given using the expected number of level-zero upcrossings of g in the one-dimensional case. According to (Rasmussen and Williams, 2006), the mean number of level-zero upcrossings in the unit interval for a GP with squared exponential kernel is

$$\mathbb{E}(N_0) = \frac{1}{2\pi\lambda}. \tag{1}$$

As we assume $\mathbb{X}_+$ to be compact and connected and, if $\mathbb{X}_-$ is not empty, we expect the discriminative function h to have $\frac{2}{3}$ level-zero upcrossings in $\mathbb{X}$ on average (compare Figure 3). If we scale $\mathbb{X}$ to the unit interval, this results in a length-scale of

$$\lambda = \frac{\Delta x}{2\pi\mathbb{E}(N_0)} = \frac{\Delta x}{2\pi\frac{2}{3}} \approx \frac{\Delta x}{4}. \tag{2}$$

The other initial hyperparameters of the classifier were set to $\sigma_g^2 = 1$ and $\tau^2 = 0.01$, based on the amplitude and the expected (low) noise level of the discriminative function. Poorly estimated hyperparameters of the regression model f have less severe consequences as those of the classifier g, as they do not result in samples outside $\mathbb{X}_+$. They will lead to a wrong density of the measurement samples' distribution, but this can be corrected by inserting additional samples, once the hyperparameters are well known. Furthermore, the limitations on the length-scales of the classifier will restrict the exploration speed and subsequently enforce a minimum sample density. Thus, no limits on

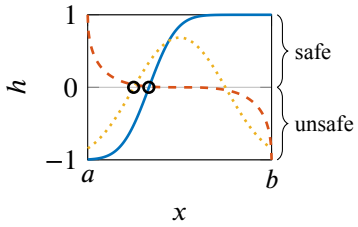


Figure 3. Three representative examples for a discriminative function $h(x)$ without noise in 1D. In this case, $\mathbb{X} = [a, b]$ and $\mathbb{X}_+ = \{x \in \mathbb{X} : h(x) > 0\}$. As one can see, two of the three examples have one zero-level upcrossing, marked with circles, the last one has none.

regression hyperparameters were enforced. The initial hyperparameters until 5 points have been sampled, were chosen as $\lambda_f = \frac{\Delta x}{4}$, $\sigma_f^2 = \left(\frac{\Delta y}{2}\right)^2$, and $\sigma^2 = \left(\frac{\Delta y}{200}\right)^2$, based on similar considerations as in the classifier case. Here, Δy denotes the expected range of the output signal y.

3.2 The discriminative model

Reference (Schreiter et al., 2015) assumes that there is only limited information about the discriminative function when measuring deep within $\mathbb{X}_+$. In contrast, the supervised output z of the HPFS system can be observed within the whole space $\mathbb{X}_+$. Thus, it is reasonable always to use the discriminative function value instead of class labels for learning the discriminative model. Another reason for doing so is hyperparameter training. Using the original algorithm and starting deep inside $\mathbb{X}_+$, it is likely that for the first samples only positive labels are drawn. It is not possible to learn correct hyperparameters from these labels using the standard training method, as shown in (Xiao et al., 2015). Therefore, the hyperparameters of the discriminative function would be estimated wrongly, which may again result in an overestimation of $\mathbb{X}_+$.

Using labels in $\mathbb{X}_-$ has some drawbacks, too. As a label only determines the discriminative function to be positive or negative in a certain point, not to have a specific value, the training algorithm can gain more information from discriminative values than from labels only. Furthermore, the regression model is not updated at all when no output y can be measured. Thus, the optima of the unconstrained part of the optimization problem described in Section 2.1 will not change if a labeled sample from $\mathbb{X}_-$ is added. In combination with the not too strong influence on the discriminative model, this may result in the next sample being generated very close to the preceding sample, as simulations showed.

For these reasons, we use a standard GP regression for the discriminative function instead of the problem specific classifier from (Schreiter et al., 2015). This has the additional advantage that no Laplace approximation for calculating the posterior is necessary. If a sample inside $\mathbb{X}_-$ is drawn, the measurement automation tries to measures at a point near the border, but inside of $\mathbb{X}_+$ instead. At this position, y as well as $\boldsymbol{z}$ can be measured. Since the original sample is usually near the boundary, the replacement sample is not far away.

3.3 The risk function

The risk function is used to encode the supervision of the systems's outputs $\boldsymbol{z}$ into a scalar discriminative function value $\tilde{h}$. This function value should be in the interval $(-1, 1)$, where -1 describes the least permissibility, 1 the highest, and 0 represents the boundary between the allowed and disallowed region.

The shape of the discriminative function h has an influence on the exploring behavior of the SAL algorithm. Comparing the continuous and the dashed line in Figure 3, one can see that the continuous line has a rather well defined zero-crossing, whereas the dashed line has only a small gradient near zero. While the former will result in a faster exploration, but with an increased risk of false positive samples in case of wrongly estimated hyperparameters, the latter will yield a slower exploration, as the discriminative model has to become very certain before samples near the boundary are queried.

By defining the risk function $\tilde{h}(\boldsymbol{z})$ accordingly, we can alter the discriminative function $h(\boldsymbol{x})$ to be learned. A proper choice could further improve the learning and exploring behavior of the algorithm. Unfortunately, $\tilde{h}$ has to be defined before starting the SAL, when the dependency of $\boldsymbol{z}$ regarding $\boldsymbol{x}$ is still unknown. Perhaps, an online optimization of $\tilde{h}$ would be possible, but this is beyond the scope of this contribution. Instead, we define $\tilde{h}$ as a linear function of $\boldsymbol{z}$. In case of the HPFS system with $\boldsymbol{z} = p$ this yields

$$\tilde{h} = 1 - \frac{p(\boldsymbol{x})}{p_{\max}} \tag{3}$$

where p is the current and $p_{\max}$ the highest allowed rail pressure.

3.4 The path to the next sample

As described in Section 2.2, we need to include the path from the GSP to the next sample in the estimation of its feasibility. Therefore, we require the constraint of the optimization problem described in Section 2.1 to be fulfilled not only at the sample point $\boldsymbol{x}_{i+1}$, but also at a number of waypoints between the GSP and $\boldsymbol{x}_{i+1}$. Even though this is only an approximation for the path's feasibility, experiments showed good results, given a sufficiently smooth discriminative function. Furthermore, it can be calculated quite fast and simple. The major drawback of this approach is that the derivative of the probability cannot be calculated analytically anymore, and thus is not available for the optimization algorithm. This results in a slightly reduced rate of convergence of the constrained optimization problem.

3.5 The algorithm

The SAL algorithm is implemented in MATLAB. Listing 1 describes the program flow in pseudocode.

Listing 1. Pseudocode for the SAL algorithm.

```
require initial measurement data D_{m_0}
    containing m_0 ≥ 1 samples, initial
    hyperparameters θ_f and θ_g, desired
    sample size n, desired safety δ
train models f and g using D_{m_0}
for i = m_0 + 1, ..., n do
        get x_i from optimization (compare
            Section 2.1)
        measure y_i and z_i and add them to
            D_{i-1} to get D_i
        calculate h(x_i) = h̃(z_i) and add it to
            D_i
        if i is large enough (compare
            Section 3.1) then
                optimize hyperparameters of
                     f and g using D_i while
                     keeping maximal
                    length-scales, where
                    applicable
        end if
        train models f and g using D_i
end for
```

4 Evaluation

4.1 Evaluation in simulation

In the first step, the SAL algorithm is evaluated using a simulated HPFS system with additive zero-mean Gaussian noise. For this purpose, a data-based GP model of the system was generated using measurement data acquired with a space-filling DoE and the ODCM algorithm presented in (Hartmann et al., 2016). ODCM allows to skip some samples based on their estimated feasibility, after a number of points in $\mathbb{X}_-$ has been sampled.

To assess the quality of the classifier, several criteria can be used. In the simulation case, we assess the classifier on a set of 10 000 test points with space-filling distribution. For the interpretation of the results, it is beneficial to consider the relative measures sensitivity and specificity. They describe the fraction of correctly classified test points in $\mathbb{X}_+$ and $\mathbb{X}_-$, respectively.

In order to benchmark the quality of the regression model, we use the root mean square error (RMSE) on m test data samples $y_{\mathrm{t},k}$ and the corresponding model outputs $f_{\mathrm{t},k}$, $k \in \{1, \ldots, m\}$.

In Table 2, a comparison of the SAL algorithm with pre- and online estimated hyperparameters is shown. In the first case, 5 to 20 initial points are sampled in a space covering about 20 % of $\mathbb{X}_+$. This region has to be defined by an expert in advance. In our case it spans about 50 % of the engine speed range and 30 % of the fuel pump actuation range. These points are used for learning the hyperparameters, which are not altered during the following SAL phase. In total, 25 points are sampled and used for modeling. At the end, the hyperparameters are optimized again using all available training data. In case of SAL without predetermined hyperparameters, these are learned online using the

Table 2. Comparison of SAL with pre- and online-learned hyperparameters in simulation after 25 samples

	pre-learned			online-learned
Initial points	5	12	20	0
RMSE	0.5900	0.0383	0.1340	0.0386
Sensitivity	0.7845	0.9928	0.9785	0.9941
Specificity	0.9647	1	1	1
Samples in $\mathbb{X}_-$	14.7	1.2	0.1	0.8

approach described in Section 3.1. All values are averaged over ten runs of the algorithm.

As one can see, 5 initial points are not enough for proper hyperparameter optimization. This leads to a disadvantageous placement of the samples, many samples in $\mathbb{X}_-$, and, subsequently, to a bad model with high RMSE. For the initial hyperparameter estimation, 12 points are a better choice, since this leads to the lowest RMSE of the final model and a small number of samples in $\mathbb{X}_-$. Nonetheless, we can outperform this variant regarding the number of unwanted samples with our online hyperparameter learning approach. The number of unwanted samples can be further reduced using 20 initial points. The remaining 5 points after the initialization are not enough to explore the whole $\mathbb{X}_+$ though, which results in a worse RMSE and reduced sensitivity compared to 12 initial points and the online learning approach.

Despite that, the variants using initial points show another unwanted property, which is not obvious from the data in Table 2: The first points during the SAL phase are often sampled far away from the initial space. In case the initial space is representative for the whole input space, this might indicate a well trained discriminative model which is able to extrapolate. This holds true in case of the HPFS system. Nonetheless, this step induces a high risk of samples in $\mathbb{X}_-$ if the initial space is not exactly representative and the discriminative function increases faster than estimated on the outside. In the online learning case, we do not observe such behavior, but a more steady exploration.

The sensitivity can be seen as a measure for the coverage of $\mathbb{X}_+$. One can see the same pattern as with the other quality parameters: The online learning approach performs best, closely followed by the 12 initial points-case.

With the right number of initial points, using pre-estimated hyperparameters performs almost equally well as the online hyperparameter learning variant. Nevertheless, it shows several drawbacks: The need to define a safe subspace in advance using expert knowledge, the right number of initial points that is hard to choose, and the large steps outside the initial space once the SAL algorithm is started. All of these drawbacks can be overcome using the online hyperparameter learning approach.

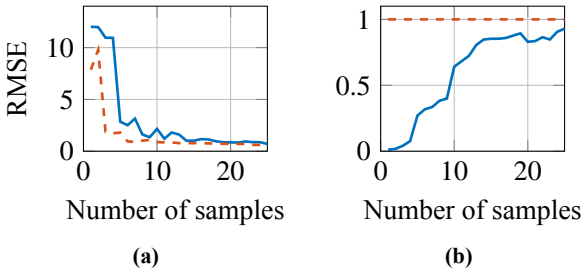


Figure 4. (a) shows the RMSE using SAL (continuous line) and space-filling plans (dashed line) at a test vehicle. The RMSE was calculated using a set of 60 space-fillingly distributed test points in $\mathbb{X}_+$. After 25 measured samples, the SAL approach obtains a RMSE of 0.7045, while a space-filling DoE results in 0.6225. (b) displays sensitivity (continuous line) and specificity (dashed line) using SAL at the test vehicle. The measures were calculated using a set of 129 space-filling test points in $\mathbb{X}$. After 25 samples, the sensitivity reaches 0.9294 and does not change considerably with a further increasing number of samples.

4.2 Evaluation at a test vehicle

After successful test runs in simulation, we apply the SAL algorithm with online hyperparameter estimation to the HPFS system of a real test vehicle. For this purpose, we use a car with 1.4 L four-cylinder gasoline engine. We implemented an automation in MATLAB, which is able to read and write labels from and to the ECU via ETAS INCA MIP, ETAS INCA, and an ETK (an ECU interface).

In Figure 4a, a comparison of the RMSEs achieved by SAL and space-filling designs over the number of training points is given. Our approach performs comparably to the space-filling DoEs. This seems acceptable, as the space-filling DoEs do not consider any limits in the input space and cannot be conducted easily this way. More measurement points do not result in a major improvement of the RMSE. The measured points in input space are shown in Figure 5.

Figure 4b shows sensitivity and specificity for each step of the SAL algorithm averaged over two runs. As one can see, the sensitivity rises rather continuously until most of $\mathbb{X}_+$ is explored. The specificity stays at its maximum value of 1 for the whole time. Note that the number of test points is much smaller compared to the simulation case, which leads to a decreased resolution of sensitivity and specificity. In average over two runs of the SAL algorithm, 0.5 points in $\mathbb{X}_-$ are sampled. As all unwanted points are sampled near the boundary, only little risk for the engine arises from them.

5 Conclusions

The test runs show that the introduced variant of the SAL approach manages to obtain a good model of the HPFS system while correctly estimating the limits of the drivable region $\mathbb{X}_+$. Only very few points are sampled in $\mathbb{X}_-$ and those which are, lie near the boundary. The resulting

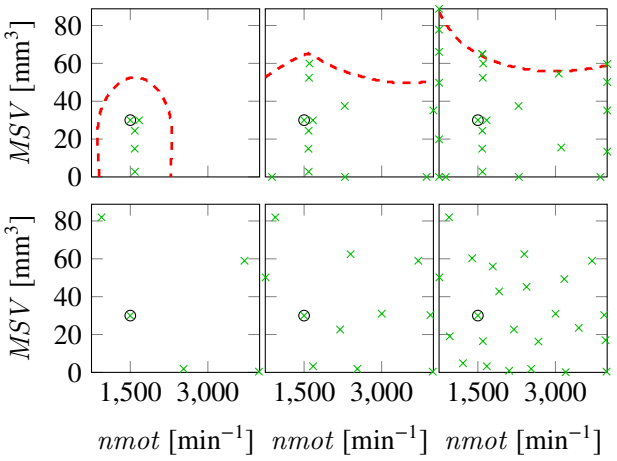


Figure 5. Measured points (crosses) in case of SAL (upper plots) and space-filling DoEs (lower plots). The algorithmic steps and plans are shown for 5, 12, and 25 points, respectively. The GSP is indicated by a circle. In the SAL case, the current estimated boundary $\mathbb{X}_0$ is plotted as dashed line.

model is almost as good as a model learned from space-filling distributed data. It must be pointed out that the latter does not comply with safety constraints or implement an exploration scheme, in contrast to the SAL approach.

References

B. Hartmann, E. Kloppenburg, P. Heuser, and R. Diener. Online-methods for engine test bed measurements considering engine limits. In *16th Stuttgart International Symposium*, Wiesbaden, 2016. Springer Fachmedien. doi:10.1007/978-3-658-13255-2_92.

C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. The MIT Press, 2006. ISBN 026218253X.

Robert Bosch GmbH, editor. *Ottomotor-Management. Systeme und Komponenten.* Friedrich Vieweg & Sohn Verlag, 3rd edition, 2005. ISBN 3-8348-0037-6.

J. Schreiter, D. Nguyen-Tuong, M. Eberts, B. Bischoff, H. Markert, and M. Toussaint. Safe exploration for active learning with gaussian processes. In *Machine Learning and Knowledge Discovery in Databases*, pages 133–149. Springer, 2015.

B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

N. Tietze, U. Konigorski, C. Fleck, and D. Nguyen-Tuong. Model-based calibration of engine controller using automated transient design of experiment. In *14th Stuttgart International Symposium*, Wiesbaden, 2014. Springer Fachmedien. doi:10.1007/978-3-658-05130-3_111.

Y. Xiao, H. Wang, and W. Xu. Hyperparameter selection for gaussian process one-class classification. *Neural Networks and Learning Systems, IEEE Transactions on*, 26(9):2182–2187, 2015. ISSN 2162-237X. doi:10.1109/TNNLS.2014.2363457.

Make Space!: Disruption Analysis of the A380 Operation in Mexico City Airport

Miguel Mujica Mota[1] Catya Zuniga[2] Geert Boosten[3]

[1,3]Aviation Academy, Amsterdam U. of Applied Sciences, The Netherlands,
{m.mujica.mota,g.boosten}@hva.nl
[2]National Aviation U. of Queretaro, Mexico, {catya.zuniga}@unaq.edu.mx

Abstract

Recently, the super heavy aircraft A380 started operations between Mexico City and Paris, and it has been announced daily operations in March. In addition, Lufthansa and Emirates are also willing to use the A380 to operate from Frankfurt and Dubai to Mexico, respectively. However, in recent years, Mexico City International Airport has been reporting severe congestion problems and it is a concern whether these problems can be overcome with the current facilities and procedures together with the increasing aircraft demand. In this article, a capacity analysis of the operation performed in the airport is presented using information for a particular high-season day. A model-based approach which allows simulating the daily operation of the A380 is presented. This approach allows incorporating most of the restrictions besides the stochasticity inherent to the system.

Keywords: A380, Mexico City Airport, simulation, performance

1 Introduction

Recently the A380 started operations between Mexico City and Paris. Some operative problems have been reported and it is a concern whether these problems can be overcome with the current facilities. To make the situation even more challenging, Air France will start daily operations in March or April and also Lufthansa and Emirates are willing to use the A380 to operate from Frankfurt and Dubai respectively (CAPA, 2014). The situation represents a challenge for the airport authorities, first accommodating the non-stopping increasing demand and second the change in the aircraft mix will affect the overall airport performance since the A380 has special requirements (Airbus, 2011). In addition, airport authorities have declared the airport congested in different slots in the last years claiming that the traffic is that high that the airport is not able to cope with the demand reported (AICM, 2015).

For assessing the current situation, we developed a stochastic model of the current airport in which the most relevant technical restrictions are incorporated. In addition, it is also implemented the current operation and routes followed by the aircraft in the ground once it lands.

2 The importance of Mexico City Airport

The Mexican air transport system transported in 2015 over 65 million passengers, an increase of 8.5% compared with the previous year. The total number of operations reached more than 1 million in that year, 748,000 of the total corresponded to national flights and 281,000 to international ones (FlightStats, 2015). This growth has supported the employment of 56.6 million people (direct and indirect jobs) and contributed with over 2.2 trillion USD to global GDP. On the other hand, the domestic sector has been growing as fast as the international one; it increased by 10% over the previous year transporting 34 million passengers (60% of the total) while the international increased a 7% moving 22 million passengers (SCT, 2015).

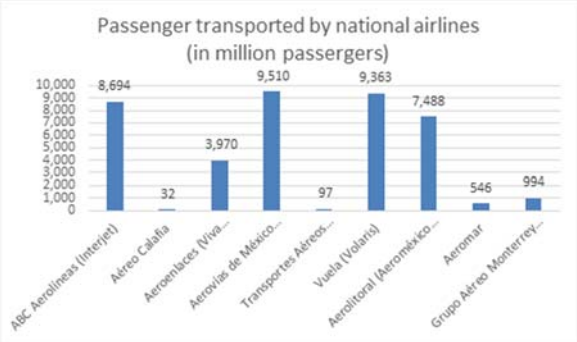


Figure 1. Passengers transported by national airlines in domestic and international routes in 2014.

Figure 1 shows the demand of the 9 regular passenger commercial airlines in México which served domestic and international routes in 2014. It can be noticed that the biggest national airlines in terms of transported passengers are Aeromexico, Volaris, Interjet and Aeromexico-Connect which moved 9.5, 9.3, 8.7 and 7.5 million pax respectively. The rest of passengers (298 000), were transported by 8 charter airlines (SCT, 2015).

Regarding to low-cost carriers (LCC), Viva Aerobus, which started operations in 2006 is growing quite fast and it is forecasted to be one of the leaders in the low-cost sector. In fact, as it can be seen in Figure 2, the low-cost sector has been growing since 2005, and in 2013 it already accounted with 60% of the market. Volaris and Interjet together with Viva Aerobus are categorized as the current Mexican low-cost carriers.

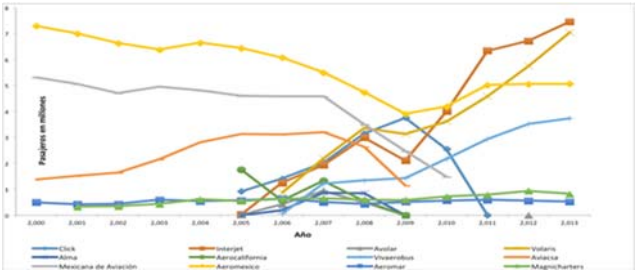


Figure 2. Main development of mexican airlines since 2005.

Table 1 introduces the top 10 domestic routes; from those routes, 47 concentrate 80.2% of the total passengers; while 80% of the international travelers use 94 routes and the 10 most frequent are presented in Table 2.

Table 1. Top 10 Domestic Routes in Mexico.

	Origin	Destination	Transported passengers (thousands)		Growing	Origin-Destination vs. Total %	
			2013	2014	2013/2014	2013	2014
1	Mexico	Cancun	3,295	3,524	7.0%	10.8%	10.7%
2	Monterrey	Mexico	2,460	2,736	11.2%	8.1%	8.3%
3	Mexico	Guadalajara	2,278	2,379	4.4%	7.5%	7.2%
4	Tijuana	Mexico	1,241	1,266	2.0%	4.1%	3.8%
5	Mexico	Merida	1,050	1,131	7.8%	3.4%	3.4%
6	Tijuana	Guadalajara	941	1,025	9.0%	3.1%	3.1%
7	Villahermosa	Mexico	700	776	11.0%	2.3%	2.4%
8	Tuxtla Gutierrez	Mexico	684	728	6.5%	2.2%	2.2%
9	Monterrey	Cancun	673	712	5.9%	2.2%	2.2%
10	Puerto Vallarta	Mexico	527	606	14.9%	1.7%	1.8%

Table 2. Top 10 International Routes in Mexico.

	Origin	Destination	Transported passengers (thousands)		Growing	Origin-Destination vs. Total %	
			2013	2014	2013/2014	2013	2014
1	Mexico	Los Angeles	783	813	3.8%	2.7%	2.5%
2	New York	Cancun	731	803	9.8%	2.5%	2.5%
3	Los Angeles	Guadalajara	746	781	4.7%	2.5%	2.4%
4	New York	Mexico	710	760	7.2%	2.4%	2.4%
5	Cancun	Atlanta	661	704	6.6%	2.2%	2.2%
6	Miami	Mexico	718	694	-3.4%	2.4%	2.2%
7	Mexico	Houston	620	693	11.7%	2.1%	2.1%
8	Dallas	Cancun	630	678	7.7%	2.1%	2.1%
9	Houston	Cancun	561	585	4.3%	1.9%	1.8%
10	Mexico	Bogota	469	572	21.9%	1.6%	1.8%

Mexico counts with 76 airports, 63 of them are international airports and 13 national; in addition there are 1,431 aerodromes registered in the country. This places Mexico as one of the top countries in Latin America with the major airport network. Figure 3 presents the 10 top airports by passenger traffic within Mexico in 2015. It can be noticed, that Mexico City International airport moves the 35% the total domestic traffic of the country, followed by four other airports: Monterrey (10%), Guadalajara (9%), Cancun (8%) and Tijuana (6%), respectively. In the international context, Cancun International airport is a good opponent to Mexico City airport moving 34% and 33% of the total, respectively.

It can be said that the busiest airport in the country is Mexico City International Airport (ICAO code: MMMX), located in Mexico city, and which also conforms, since 2003 the pillar of the Metropolitan Airport system, together with Queretaro, Puebla, Toluca and Cuernavaca.

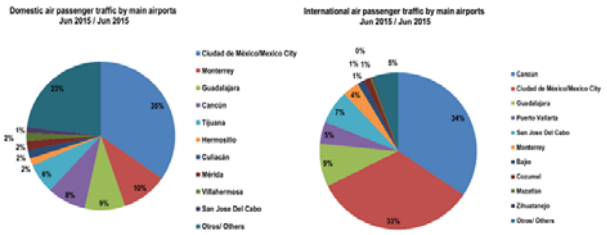


Figure 3. Domestic and International passenger traffic by main airports in Mexico.

Regarding Mexico City Airport, it is considered key for the development of the metropolitan region in Mexico and also for the development of the country. Recently it has been announced the development of the new airport in Mexico City which will have a final capacity of 120 mill pax/yr (Herald On Line, 2014). However the first phase for this airport will not be operative until 2020. In the meantime, Mexico City as a destination is still growing and the country has also gained importance as a tourist and business destination. On the 12th of January 2016, AirFrance started a direct flight from Paris to Mexico City using the mega jumbo A380. At the moment, the flight is only scheduled 3 times a week but it is planned that from March on it will fly on a daily basis (Experience the Skies, 2016). Each flight of the mega Jumbo transports 516 passengers and due to its dimensions and operational requirements some problems have raised in which delays are the most relevant ones.

The flight to and from Paris represents itself a challenge to the airport due to different factors which cause problems to be solved by the airport operator. One problem is that the clearances from the centerline at some taxiways are too narrow for the size of the aircraft which has caused that the aircraft follows a long taxi route to the runway (SENEAMM API, 2015). This operative situation caused that the departure time suffers a delay of 10 to 56 minutes with an average value of 36 minutes (Experience the Skies, 2016). On top of this situation, some years ago the airport authorities

established a limit of 61 ATM/HR as the maximum hour capacity for the airport, hence some slots of the airport have been declared congested. Furthermore, Lufthansa and Emirates have stated that they have intentions to start operating with the A380 from Frankfurt and Dubai to Mexico city respectively. For these reasons is critical to have tools and methodologies that allow the study the current and future operation of the airport. Traditional analytical techniques have proven their lack of power for addressing with accuracy the potential problems of a complex system such as the one of Mexico City airport, that is the reason model-based techniques and in particular stochastic modeling appears as the only one with the capability for analyzing with the proper accuracy the current and future operative situations.

Initial simulation-based approaches have been performed by different authors addressing the problems in MMMX and other airports (Herrera, 2012; Bazargan, 2004; Marelli et al, 1998; Soolaki et al, 2012; Mujica, 2015). In this work, the analysis performed using a validated model of the current operation of the Airport of Mexico City it is presented which allows the understanding of the current potential problems and also the ones that will rise once the daily operation of AirFrance takes place.

3 Model-based Approach

The developed model is a discrete-event-based model which allows including the stochastic characteristics and level of detail that other analytical approaches would not allow. The level of detail is such, that enables the integration of the technical restrictions, the operative restrictions imposed by the airport authority, the rules in place for the different aircraft such as wake-vortex separation and the taxiway routing for landing and takeoff. The elements that compose the complete model are: the two runways, taxi network, terminal buildings, and parking stands of the two terminals. The model focuses only in the airside of the airport and it does not pay attention to the airspace, flow of passengers or vehicles that perform the services within it, so it is bounded to the airside operation only. Figure 4 illustrates the layout of the model that includes the taxiway network, airport stands and runways and it also depicts the different paths that are followed by the traffic within the airside.

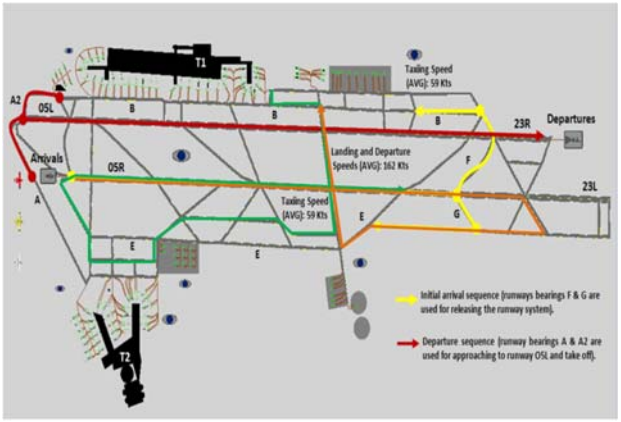


Figure 4. MMMX airport layout with A380.

The yellow path illustrates the normal landing configuration and the red line represents the configuration followed for departing flights. However, the situation of the A380 is slightly different; the A380 follows the orange path and green path for arrival and departure respectively (SCT, 2015).

In order to make the model valid, different characteristics were included in the model besides different assumptions. The most relevant ones are presented in Table 3.

Table 3. Characteristics of the Airside Model.

Parameter	Value
Landing Speed	Min:150 Knot, Max 175 Knot, AVG 162 Knot
Taxiing Speed	Min: 49 Knot, Max: 68 Knot, Avg: 59 Knot
RWY O5L-23R	Length: 3963 m
RWY 05 R-23L	Length: 3985 m
Number of Stands	T1: 50, T2:46
CenterLine Separation	310 m
Turnaround Time	Probability Distributions depending on the type of AC

For the traffic demand generation of the model, information from a representative day was collected. The information was considered (FlightStats, 2015), (Flight Radar24, 2015) and then the performance of the model was compared against the real number of air transport movements of the day.

In order to evaluate the impact of the A380,information from the current operation was collected, the type of information that was included in the model is the following:

- Route of Taxi-In and Taxi-Out of the A380
- Speed of the Taxi-In/out of the A380 in the Airport
- Turnaround time
- Current schedule and gate allocation

The operation of the airport has been modified in order to cope with the challenge of giving space for the A380 to operate. Due to the limitations and restriction in the operation, the route of the aircraft is not the standard one but a modified one so that the aircraft is

able to get to the gates G33 and G34 which were the ones selected for the operation of the A380.

3.1 Experiments and Analysis

The results were obtained running first the case without the A380 in order to make possible to establish a base case for comparison. Once the results were obtained with the base case, modifications to the model were made and attention was paid to some performance indicators.

Scenario 1: Base Case

First the base model was run and the utilization of the gates and the number of operations per hour (ATM/HR) was obtained from the initial replications, Figure 5 illustrates the values of these indicators.

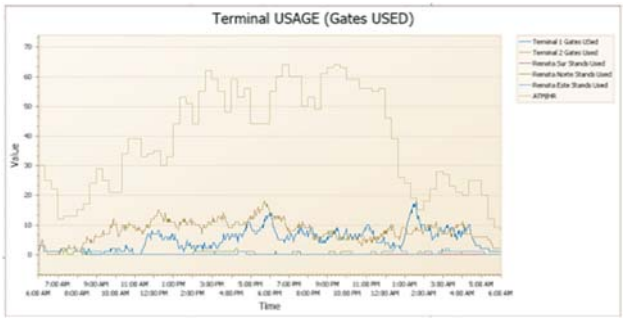


Figure 5. Gate usage and ATM/hr during the day.

Based on the analysis of the base case, it was identified an unbalance in the gates throughout the day. From Figure 5, it can be seen that most of the time the usage of Terminal 2 (brown line) is higher than the usage of Terminal 1 (blue line). It can also be appreciated that, as it has been claimed, during some time of the day, the number of operations has exceeded the declared capacity of the airport (61 Atm/hr), in particular form 1:00 pm until 10:00 pm.

In order to understand the performance capacities of the airport, a statistical analysis of the system was performed and it was evaluated whether the use of the A380 have impacted the operation. For the base case, 30 replications were performed and characteristics of the system were obtained. The most important results are presented in Table 4. From the initial results it can be appreciated that the traffic is apparently unbalanced since the Terminal 2 uses more gates in average than the Terminal 1 during the day. The maximum values represent that at some period of time during the day approximately 21 gates out of the 34 are used. In addition it can also be appreciated that the runway usage is about 69-70% as it can be seen in the green line of Figure 6.

Table 4. Base Model without A380.

Facility	Gate utilization	NO- A380		
	Max Avg	Min	Max	HW
Terminal 1 (36 Gates)	14.4	10	28	1.33
Terminal 2 (34 Gates)	15.2	11	21	0.88
	AVG	Min	Max	HW
Ratio T1Gates/T 2 Gates	0.43	0.34	0.57	0.02
TWY T1 queue	5.5	4	8	0.4
TWY T2, Queue	3.8	2	5	0.33

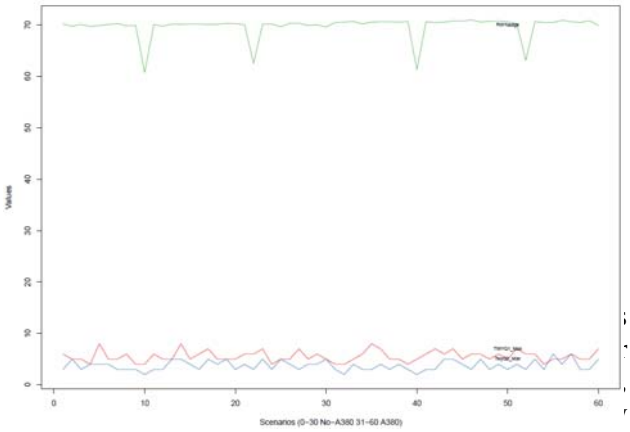


Figure 6. Utilization of runway and gates in the terminals.

This figure suggests that we can identify that the runway is the element of the system that might suffer from the lack of capacity to handle more traffic. Therefore, attention must be paid to that element.

Scenario 2: Disruptive A380

For the scenario that includes the A380, the same flight schedule was used but the flight of AirFrance at 6:40 pm was incorporated using the gates G33-G34 and following the published route for the A380 (SCT, 2015). Table 5 shows the results of scenario 2. For this scenario, 30 replications were run for obtaining the performance of the system under the new conditions.

Table 5. Scenario with A380.

Facility	Gate utilization	With- A380		
	Max Avg	Min	Max	HW
Terminal 1 (36 Gates)	14.16	10	21	0.86
Terminal 2 (34 Gates)	14.86	11	21	0.93
	AVG	Min	Max	HW
Ratio T1Gates/T 2 Gates	0.44	0.34	0.56	0.019
TWY T1 queue	5.6	4	8	0.38
TWY T2, Queue	3.7	2	6	0.39

From these results, it is apparent that in general terms the utilization rate of both terminals is affected since the maximal utilization and also the average values are negatively modified. However, it can also be appreciated that in some replications of the model and due to the variability the usage is as low as 28% and 32% for terminal 1 and terminal 2 respectively.

Impact Analysis

Once the different values and the data of the operation were obtained, *t* tests were performed for evaluating if the impact of the operation is statistically significant. These tests were executed over the following indicators with a level of significance of 0.05:

- Ratio UsageT1/UsageT2
- Effect on Queue of T1
- Effect on Runway usage
- Effect on Queue of T2
- Average T1Gate usage

After running the different tests it was possible only to identify a significant effect over the third performance indicator called *Effect on Runway Usage*. This indicator deals with the level of congestion of the runway. In our study the runway usage went from 69.5% to 70.02 %. Figure 7 illustrates the effect of the indicator once the A380 appears into scene. The dots with the 0 value correspond to the base case and the ones with the 1 value next to the dots represent the scenario with the A380.

In this figure, the effect of the inclusion of the A380 is evident since the trend is similar but just shifted to a high position in the graph due to the increase in saturation of the capacity.

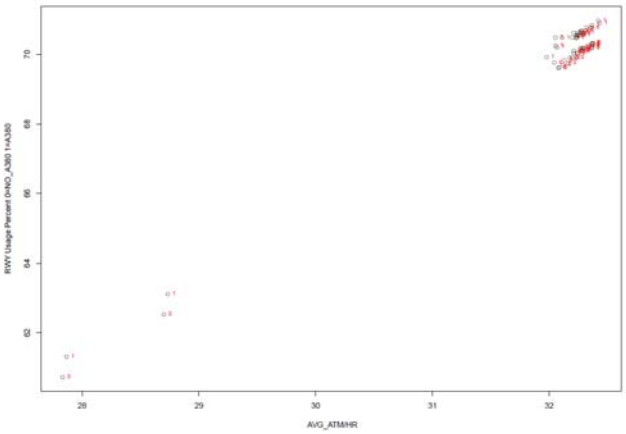


Figure 7. Effect in the RWY of the inclusion of A380.

The next figure also shows the hypothesis test performed for verifying the effect of the A380 in the runway usage. As the reader can appreciate, the shift in the mean is due to the introduction of the A380 as the *p* value confirms.

	Effect on RWY USAGE	
t-Test: Paired Two Sample for Means		
	Variable 1	*Variable 2*
Mean	69.48701814	70.02946
Variance	4.662065343	4.618695
Observations	30	30
Pearson Correlati	0.9966713	
Hypothesized Me	0	
df	29	
t Stat	-16.87623209	
P(T<=t) one-tail	7.74046E-17	
t Critical one-tail	1.699127027	
P(T<=t) two-tail	1.54809E-16	
t Critical two-tail	2.045229642	

Figure 8. Values of the Hypothesis Test.

This result confirms the statement of the airport operator that the runway is the main bottleneck in this system. However, it is also interesting that the system is able to handle this type of aircraft without affecting the operation in the remaining elements of the facility.

4 Conclusions

Airports in the globe are facilities that are very important for the development of a region in any country. Especially international airports have the function as the gateway to economic areas of development.

In the particular case of Mexico, the international airport of Mexico City is a critical infrastructure which works as a catalyst for the development of the region in different aspects that range from tourism to business. In the present article we analyzed through a model-based approach the situation of the operation of the A380 which recently started operations to and from Paris.

We could identify peculiarities of the operation of the airport such as the unbalance of the use of the gates in Terminal 1 and Terminal 2. We run experiments over a couple of scenarios and we could also identify that the operation of the A380 only affected the runway within the system. This result is interesting since it demonstrates that the A380 is not as disruptive as it was initially expected. However, the runway is the most sensitive echelon in the system and the airport should put more focus in the runway management if they want to allow other carriers such as Emirates or Lufthansa fly to Mexico City using A380s.

As a future work, we will study the particular situation of the peak hours of the airport in order to identify the slots that are less sensitive to the operation of a future A380.

Acknowledgements

The authors would like to thank the Aviation Academy of the University of Applied Sciences for supporting this study and the Dutch Benelux Simulation Society (www.dutchbss.org) and EUROSIM for the dissemination of the findings of this study.

References

A. Herrera. *Simulation Model of Aeronautical Operations at Congested Airports: The case of the Mexico City International Airport*. Mexican Institute of Transport, Queretaro, Mexico. Technical publication, 365, 2012.

Airbus. *A380 Airplane Characteristics for Airport Planning.* Airbus SAS, Blagnac Cedex, France. Technical Report, 326, 2011.

AICM. *Statistics and Flight-Schedules, Mexico City International Airport*. Available via https://www.aicm.com.mx [accessed September 21, 2015].

CAPA Centre for Aviation. *Mexico DGAC: Air France, Lufthansa, Emirates and Turkish Airlines interested in A380 to Mexico City*. Available via http://centreforaviation.com/news [accessed June 4, 2014].

Experience the Skies. *Airbus A380 Faces Challenges at Mexico City International Airport*. Available via http://www.experiencetheskies.com [accessed January 29, 2016].

FlightStats. *Benito Juarez International Airport Arrivals/Departures, Flight Stats Inc*. Available via https://www.flightstats.com [accessed November 18, 2015].

FlightRadar24. *Live Air Traffic*. Available via https://www.flightradar24.com/19.43,-99.1/12 [accessed December 15, 2015].

Herald on line. *Huge new airport is announced for Mexico's capital*. Available via http://www.heraldonline.com/latest-news/article12005174.html [accessed September 2, 2014].

M. Bazargan. Airline operations and Scheduling. Burlington, USA, *Ashgate Publishing Company*, 1st edition, 205, 2004.

M. Mujica. Check-In allocation improvements through the use of a Simulation-Optimization Approach. *Transportation Research Part A*, 77: 320-335, 2015.

M. Soolaki, I. Mahdavi, N. Mahdavi-Amiri, R. Hassanzadeh, and A. Aghajani. A new linear programming approach and genetic algorithm for solving airline boarding problem. *Applied Mathematical Modelling*, 36 (9): 4060-4072, 2012.

S. Marelli, G. Mattocks, and R. Merry. AERO Magazine 1. *The Role of Computer Simulation in Reducing Airplane Turn Time*. Available via http://www.boeing.com/commercial/aeromagazine [accessed January 6, 2014].

SCT, and SENEAMM. *eAPI, 40-MEXICO,7_AD-MMMX-2-18 Historical statistics (1992-2014)*. Secretaria de Comunicaciones y Transportes (SCT), Mexico City, Mexico. Technical report, 2015.

A Causal Model for Air Traffic Analysis Considering Induced Collision Scenarios

Marko Radanovic and Miquel Angel Piera Eroles

Department of Telecommunications and Systems Engineering,
Autonomous University of Barcelona, Spain,
{marko.radanovic, miquelangel.piera}@uab.cat

Abstract

Present research on the air traffic management systems is trying to improve an airspace capacity, accessibility and cost-efficiency while maintaining the safety performance indicators. The discretization of the aircraft trajectories in a sequence of the 4D points specifying an agreement between the airspace users and the traffic flow management, in which the aircraft are required to arrive at the certain waypoints in the required time instants, opens a huge scope of applications for the decision support tools. This paper presents the causal model of an induced collision scenario, generated by the Traffic alert and Collision Avoidance System logic, tailored by an impropriate pairwise collision resolutions. It elaborates a unit simulation case and introduces a new modeling approach through the Colored Petri Net formalism. The proposed model provides a better insight on the geometry of collision trajectories which is a baseline for the simulation of new conflict-free resolution strategies that could be automated, and integrated in the further research.

Keywords: causal modeling, hotspot, induced collision, pairwise encounter, resolution advisories

1 Introduction

The constant increase in the air transport demand is generating a continuous pressure on the air traffic control (ATC) system. As a result, more efforts in the ATC modernization have been made to satisfy the main ATM criteria: enhanced capacity, cost-efficiency and safety. Based on the Single European Sky ATM Research (SESAR) initiative (Drogoul *et al*, 2009), there would be necessary to shift from the completely centralized tactical ATC interventions to more efficient, decentralized, collision-avoidance operations. This foresees the important changes in the roles and responsibilities of the overall air traffic management (ATM) system.

At present, an upgraded Traffic Alert and Collision Avoidance System (TCAS II v7.1), has been designed for operations in the traffic densities of 0.3 aircraft per squared nautical mile. The system demonstrates an excellent performance in cases of the pairwise encounters (PEs) but, concurrently shows some performance drawbacks in its logic due to the well reported induced collisions in some traffic scenarios (Jun *et al*, 2014, 2015; Ruiz *et al*, 2013). These drawbacks are also a result of frequent changes in the kinematic trajectory elements (the speed and altitude changes), as well as an ambiguity in the horizontal level crossings and level busts. Thus, one of the goals will be to investigate and implement a new operational framework improving the TCAS functionalities to react at both tactical and operational level as a robust collision avoidance system for different complexities of the traffic scenarios, in which ergonomics and automation interdependencies will be fully considered and aligned with the realistic aircraft performances.

This paper analyzes a pairwise collision scenario as a product of the previously resolved conflicts. From a causal point of view, it illustrates the case in which some inappropriate maneuvers, issued by TCAS to solve one-on-one encounters, can induce a new collision. This effect is known as a downstream effect (i.e. emergent dynamics) of the previous TCAS decisions and can be treated as a surrounding traffic effect, separately from a multi-threat encounter approach. The scenario is then simulated using an open-source conflict detection and resolution (CD&R) toolset *Stratway*, and obtained results are presented. Based on the simulated case, a new approach is introduced through development and validation of a Colored Petri Net (CPN) model. The model presents a baseline for further research on a probabilistic, state-based collision prediction.

The remainder of the paper is organized as follows. Section 2 discusses a conceptual analysis of the hotspot scenario describing both the one-on-one encounters and pairwise induced collisions. A simulated scenario with the obtained results is presented in Section 3, while Section 4 describes the causal modeling approach for the collision prediction. Section 5 validates the presented model, and conclusions are given in Section 6.

2 Conceptual Analysis

A reduction of the separation minima might occur due to many circumstances; in a moment when an air traffic controller issues a resolution directive, in which any change in a desired cruising speed, heading or vertical rate is not appropriate, or when a pilot performs a

maneuver that a controller had not anticipated. The airspace volume that encompasses a subset of trajectories with tight spatiotemporal interdependencies, which can easily lead to reduction of the separation minima, defines a hotspot. This volume is both space- and time-dependent on the aircraft closure rates, in sense that can occupy a couple of flight levels (several thousands of feet's) and a longer horizontal distance (several tens of nautical miles).

An idea of the PE approach lies in fact that an induced collision with the closest points of approach (CPA) of two aircraft cannot dimension the hotspot area, which is not in the case of the multi-threat encounters. Instead, a surrounding traffic aircraft introduce a certain level of uncertainty in the geometry of a resolution trajectories and, thus, very tight spatiotemporal interdependencies between trajectories that can be involved in collision are essential to define the hotspot itself (Billingsley *et al*, 2013). The CPA is an estimated 4D point on the aircraft trajectory, for which a 3D distance between two conflicting aircraft reaches its minimum value.

Even if assumed that integrity levels of the 4D trajectories are fully accomplished (i.e. very small along-track, across-track and vertical path deviations) and the flight parameters (heading, altitude and speed) are progressively maintained, which also imply the constant timestamp changes, it is not possible to predict an induced CPA. Naturally, this question opens many analytical aspects, but the main ones are a limited TCAS logic, based on a certain number of the resolution advisories (RAs), TCAS threshold requirements, and the feasible maneuvering strategies based on the aircraft performance (ICAO, 2006).

Table 1 lists all advisories for TCAS II v 7.1, while Table 2 depicts the TCAS threshold values for different flight levels. The second column in Table 2 refers to the sensitivity level (SL) indexes. This one-digit number features a strength sense of TCAS command.

Table 1. TCAS Advisories.

TCAS II v 7.1			
Type	*Text*	*Meaning*	*Required Action*
TA	Traffic, traffic	Intruder near both horizontally and vertically	Attempt visual contact, and be prepared to maneuver if RA occurs
RA	Climb, climb	Intruder will pass below	Begin climbing at 1500-2000 ft/min
RA	Descend, descend	Intruder will pass above	Begin descending at 1500-2000 ft/min
RA	Increase climb	Intruder will pass just below	Climb at 2500-3000 ft/min
RA	Increase descent	Intruder will pass just above	Descend at 2500-3000 ft/min
RA	Reduce climb	Intruder is probably well below	Climb at slower rate
RA	Reduce descent	Intruder is probably well above	Descend at slower rate
RA	Climb, climb now	Intruder that was passing above, will now pass below	Change from descent to climb[1]
RA	Descend, descend now	Intruder that was passing below, will now pass above	Change from climb to descent[1]
RA	Maintain vertical speed, maintain	Intruder will be avoided if vertical rate is maintained	Maintain current vertical rate
RA	Level off, level off	Intruder considerably away, or weakening of initial RA	Begin to level off
RA	Monitor vertical speed	Intruder ahead in level flight, above or below	Remain in level flight
RA	Crossing	Passing through intruder's level, usually added to any other RA	Proceed according to associated RA
CC	Clear of conflict	Intruder is no longer a threat	Return promptly to previous ATC clearance

[1]This is reversal RA that requires change of 1500 ft/min vertical rate.

Table 2. TCAS Threshold Values.

Own Altitude [ft]	*SL*	*TAU [sec]*		*DMOD [NM]*		*ZTHR [ft]*		*ALIM [ft]*
		TA	*RA*	*TA*	*RA*	*TA*	*RA*	*RA*
1000 - 2350	3	25	15	0.33	0.20	850	600	300
2350 - 5000	4	30	20	0.48	0.35	850	600	300
5000 - 10000	5	40	25	0.75	0.55	850	600	350
10000 - 20000	6	45	30	1.00	0.80	850	600	400
20000 - 42000	7	48	35	1.30	1.10	850	700	600
> 42000	7	48	35	1.30	1.10	1200	800	700

2.1 CD&R for One-On-One Encounters

To explain the concept of induced collision, it is first considered an initial state of a non-vectored traffic scenario in the vertical plane, which presents the SESAR concept for a free routing without a level capping. There are four aircraft A/C01, A/C02, A/C03 and A/C04 flying on the trajectories that form two predicted encounters A/C01-A/C02 and A/C03-A/C04 (Figure 1).

A/C01 is cruising on FL160 while A/C02 starts descending at FL180 in the opposite direction from A/C01, which means a direct approch to A/C01 with a loss of height. On the other side, A/C03 starts climbing at FL130, and, with an increase in height, approaching to A/C04, which is crusing at FL153 in opposite direction from A/C01. The sequences of 4D waypoints (WPs) for all four trajectories are assumed to be charaterized by the same absolute timestamps, which confirms the time-based dimension of potential hotspot. It can be noted that a difference in altitude between A/C01 and A/C04 is only 700 feet, and in this case TCAS vertical threshold is still satisfied since the hotspot area belongs to SL6 with an RA activation at the 600-feet difference. Therefore, by concluded that these two aircraft operationally maintain the required vertical separation.

As known, in the normal flight conditions TCAS is incessantly surveying the surrounding airspace by sending queries (interrogations) and receiving responses from the neighbouring aircraft. Therefore, when A/C02 flies into the range of A/C01, the TCAS on-board both aircraft issues traffic advisory (TA) to warn the crew about a possible conflict. In this scenario, the TCAS advises A/C01 and A/C02 in moments t_{TA}^{01} and t_{TA}^{02}, respectively. Naturally, this warning is activated if and only if all three TCAS thresholds for the particular SL are infringed (Table 2). Based on the current flight configuration of both aircraft and approaching closer to each other, at the instances t_{RA1}^{01} and t_{RA}^{02} TCAS issues the RAs requiring that both aircraft perfom an appropriate maneuver (Table 1). Moreover, the RAs are also a subject to the TCAS threshold infringements (Table 2).

The corresponding maneuver depends on the CPA which is determind by speed, heading and position of the aircraft. It is worth mentioning that, due to high level of range-bearing errors in the horizontal plane (Kochenderfer *et al*, 2013), the RAs consider only maneuvers in vertical plane, possibly combined with some turns or heading changes. However, those cases are a measure of the aircraft performances and the crew experiences, and are out of scope of this study. There are four rules in the TCAS logic for the PEs:

1. Two aircraft are alerted by the RAs when the horizontal and vertical threshold distances, DMOD and ZTHR respectively, are violated, or when the time to the CPA (TAU) falls below a specific threshold, with respect to the current aircraft closure rates and a corresponding SL.

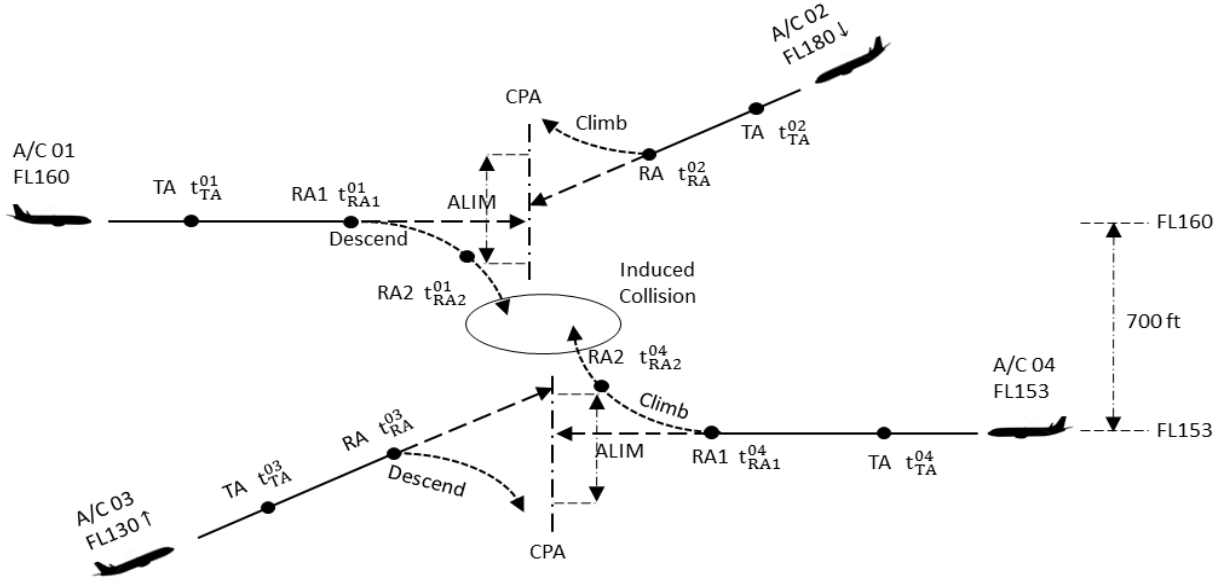


Figure 1. An induced collision scenario.

2. Two RAs are opposite to each other, i.e. they advise an opposite sense for maneuver to the crew (for instance, climb-descend or descend-climb). It is defined as a *reversal* TCAS logic.
3. When the RAs are alerted, an aircraft at a lower altitude performs descending maneuver and the one at a higher altitude complies to a climbing amendment, without consideration of the current flight configuration (cruise, climb or descent); However, the strength sense of the requested manouver will depend on the flight configuration consequently (Table 1).
4. After the RA activation the aircraft following the requested amendments must achieve a vertical separation minima at the CPA, called the altitude limitation (ALIM), as illustrated in Figure 2.

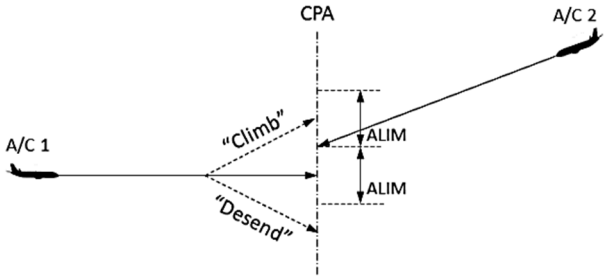


Figure 2. "Descend" RA to achieve ALIM.

TCAS computes TAU as a ratio between the range (an interdistance among the aircraft) and closure rate (or, range rate). Both range and range rate in horizontal plane are obtained from the TCAS interrogations, usually with one-second update, and they apply to aircraft in crusing configuration. In vertical plane, the time to co-altitude (vertical TAU) is computed as a vertical separation divided by a vertical closure rate (Jun *et al*, 2015).

It can be observed from Figure 1 that A/C01 at moment t_{RA1}^{01}, from the crusing phase, starts descending while A/C02 changes from descending to climbing manouver at t_{RA}^{02}, both achieving required separation minima at the CPA. Another PE, A/C03-A/C04, results in a similar way. The TAs in the moments t_{TA}^{03} and t_{TA}^{04} warn A/C03 and A/C04 about a potential conflict. At t_{RA}^{03} and t_{RA1}^{04}, the RAs are activated and both aircraft start performing the advised maneuvers. A/C03 is passing to the descending and A/C04 to the climbing amendment. In practice, it could happen that any of the Ras is not properly applied due to some unpredictable factor(s) (a meteo situation – a wind component, lack of the requested aircraft performance, or any technical error on-board the aircraft). In this case, the ALIM might be infringed and the conflict evolves into a collision.

2.2 Induced collision scenario

The previous subsection has led to the conflict resolutions of two neighbouring encounters. However, the main question is whether these amending trajectories could possibly generate a new conflict. This induced conflict can be elaborated though the emergent dynamics concept. Based on the dimensioned hotspot, it can be observed that A/C02 and A/C03 leave the area on their new conflict-free paths. In other words, they achieve their clear of conflict (CC) points (Table 1).

Concurrently, by following the previous RAs A/C01 and A/C04 induce a conflict. This state could be ambiguous. If the hotspot encompasses several flight levels and a larger horizon this encounter would become an induced conflict and might remain a conflict-based with enough time for the new RAs activation. However, if there is no sufficient time, the induced collision occurs. The analyzed scenario points out to that state. As a collision avoidance layer activates in less than 60 seconds and the RAs are issued in less than 35 seconds before the CPA reachability, once resolved conflicts produce very high uncertainty in guidance over the resolution amendments. Since the original trajectories of A/C01 and A/C04 have been vertically separated only by 700 ft and, by performing their resolution manouvers, the aircraft triggered the new TCAS alerts, the vertical thershold has been considerably violated. A/C01 and A/C04 were automatically alerted by the succeeding RAs, at the timestamps t_{RA2}^{01} and t_{RA2}^{04}, respectively. Due to insufficient time for the appropriate maneuvers, the aircraft came to the induced collision.

TCAS is operating in vertical plane which comprises a set of the vertical RAs only. Therefore, a collision event is predominantly affected by the upstream and downstream traffic flows.

3 Scenario Simulation – Unit Case

This section describes the simulation platform for CD&R algorithm and provides the scenario results for unit case. Results are presented both graphically (within integrated Graphical User Interface – GUI) and textually (in form of the log messages).

3.1 Simulation Platform for CD&R algorithm

For simulation of our scenario we have used the Stratway tool. It is the algorithm for a strategic, intent-based, CD&R, developed by the NASA Langley Research Center. Stratway is an open source software tool, implemented both in Java and C++ environment, and can be called from other programs through an Application Program Interface (API) and also excuted from a command line. The main features are as follows:

- work with the complete 4D flight plans as inputs (three spatial geographic coordiantes + time);

- generation of the conflict resolution in the form of conflict-free paths for the ownship aircraft (a reference trajectory) in presence of the multiple traffic aircarft (intruders), if feasible;
- use of a set of the heuristic search strategies for the conflict resolution;
- output of the message errors and warnings, as well as the textually-based solutions;
- considerably based on the real aircraft performances and use of a large set of the navigation parameters, that are user-configurable;
- implementation of a set of maneuvering strategies (*vertical*, *track*, *speed*, or *side-step*) which are 3D-oriented (Figures 3, 4 and 5);
- iterative tests of all involved trajectories, and output of all possible combinations for the trajectory resolutions;
- no current support to testing the induced collision scenarios; however, there are possibilities for making the upgrades with the new functionalities and strategies.

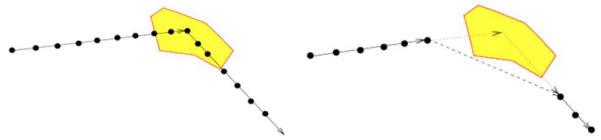


Figure 3. Track search strategy.

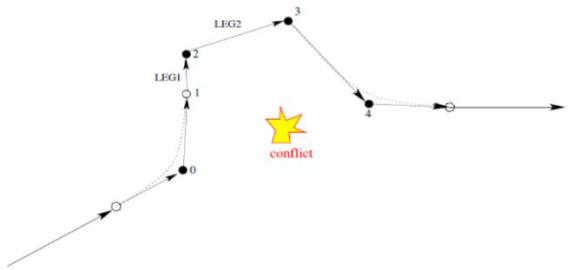


Figure 4. Vertical search strategy.

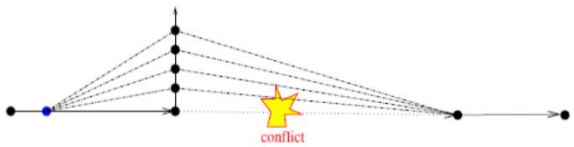


Figure 5. Side-step search strategy.

Figures above illustrate three types of maneuvering strategies. The first type is a *track* strategy (Figure 3) which is based on a heading change. The goal is to avoid a hotspot by isolating the WPs, positioned inside the hotspot, and to directly re-route to the first available WP outside the hotspot. Nevertheless, this strategy can be treated as a *fly-by-waypoint* procedure, where this WP is an imaginary center of the hotspot area. The second type is a *vertical* strategy (Figure 4). It seeks to resolve the conflict through a sequence of climbs or descents without changing the current heading of the trajectory. The strategy starts from the WP in vicinity of the hotspot using the same isolation method as the track strategy. From this WP, an aircraft increases the climb amendment (or descent, in the opposite case) in order to overtop a potential intruder. The amending trajectory leg is usually shaped as a polygon consisting of the shorter segments. A *side-step* strategy (Figure 5) is the third type, and is not considered so flight-efficient as the track strategy, but sometimes can provide the comparable solutions. It resolves a conflict by only removing a WP right before the conflict. The strategy is very effective for the trajectories containing longer segments. It inserts a *lead-in* WP (blue-colored point) in advance of the current aircraft position, from which the aircraft starts with an amending leg, and then continues with a resuming leg to the original trajectory. A deviation from the original trajectory, or the point at which amending leg terminates and the resuming leg starts, depends on the geometries of the conflicting trajectories and the closure rates.

3.2 Unit Case Simulation and Results Validation

For simulation of two PEs, it has been implemented a unit case scenario within the Dortmund enroute airspace (51°30′53″ N, 7°27′57″ E), between 13000 and 18000 ft (FL130 – FL180). Each of four trajectories has been generated in a sequence of 10 WPs with the constant time-based segments (15 seconds of the time interval) in order to facilitate the encounters prediction. Closure rates, i.e. the true airspeed (TAS) in cruising and vertical speed - rate of climb/descent (ROC/D) - are assumed to be constant as well, and by default set to TAS = 330 knot and ROC/D = 1500 ft/min. Nevertheless, these values can be changed as per user preferences, or adopted to a specific SL. Table 3 illustrates a sample of the sequences of the 4D WPs for all four trajectories used as an input. *OWN* in the table denotes the ownship aircraft, while *TRAF* with the given index corresponds the traffic aircraft.

Table 3. Input Data.

Name	*Latitude [deg]*	*Longitude [deg]*	*Altitude [deg]*	*Time [sec]*
OWN	51.51389	7.53075	16000	2400
OWN	51.51389	7.55370	16000	2415
TRAF1	51.51649	7.61961	18000	2400
TRAF1	51.51604	7.61894	17625	2415
TRAF2	51.48779	7.68225	13000	2400
TRAF2	51.48824	7.68292	13375	2415
TRAF3	51.49155	7.80565	15000	2400
TRAF3	51.49155	7.78405	15000	2415

In order to graphically present the simulated results, a graphical user interface (GUI) has been developed as a part of the Stratway algorithm. The Stratway GUI is

composed of 6 views illustrating the interdependencies between 4D coordinates:

- latitude-longitude,
- altitude-longitude,
- altitude-latitude,
- altitude-time,
- latitude-time,
- longitude-time.

By default, latitude and longitude are expressed in degrees [deg], altitude in feets [ft] and time in seconds [sec]. The simulation of the Dortmund scenario in Stratway has validated that the present resolution algorithm cannot find a conflict-free path within an induced collision (Figure 6). None of four aircraft, set iteratively as the ownship, could avoid induced collision as it has occurred on the central segments of their trajectories. The Stratway also generates also the graphical output of the pairwise conflicts without a possibility for any aircraft approaching to the induced collision state to perform an appropriate RA maneuver (Figure 7).

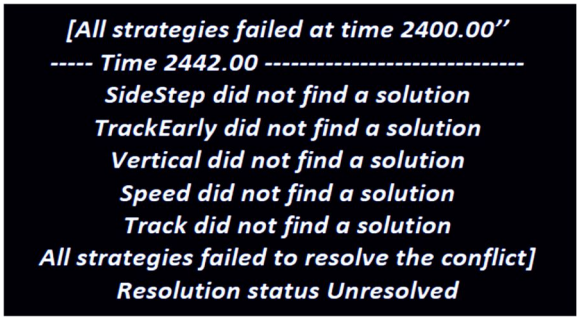


Figure 6. Log message output.

4 Causal Model for Collision Prediction

4.1 CPN Formalism

The main CPN characteristics that present very applicable formalism for a description of the discrete event-oriented simulation models are:

- all events that could appear according to a certain system state can be easily determined by a reachability graph;
- all events that can set off the *firing* of a specific event can be detected visually. CPNs are considered as a graphical modeling tool with a few syntactic rules.

The main CPN components that meet the modeling requirements are: the *places*, represented by the circles, and specifying the system states; the *transitions*, depicted by the rectangles and expressing the system events; the *input arc expressions* and *guards*, indicating the types of tokens used to fire a transition; the *output arc expressions* indicating the system state change that appears as a result of firing the transition; the *color sets*, the entity attributes which determine types, operations and functions that can be used by the elements of the CPN model; a *state vector,* the smallest piece of information for prediction of the events that could appear. This vector denotes the number of tokens in each place and the colors in each token. The color sets allow specification of the entity attributes, and the output arc expressions define what actions should be coded in the event routines linked to each event.

4.2 CPN Modeling Approach for Pairwise Collision Prediction

This subsection proposes a new causal modeling approach for the right discretization of conflict/collision

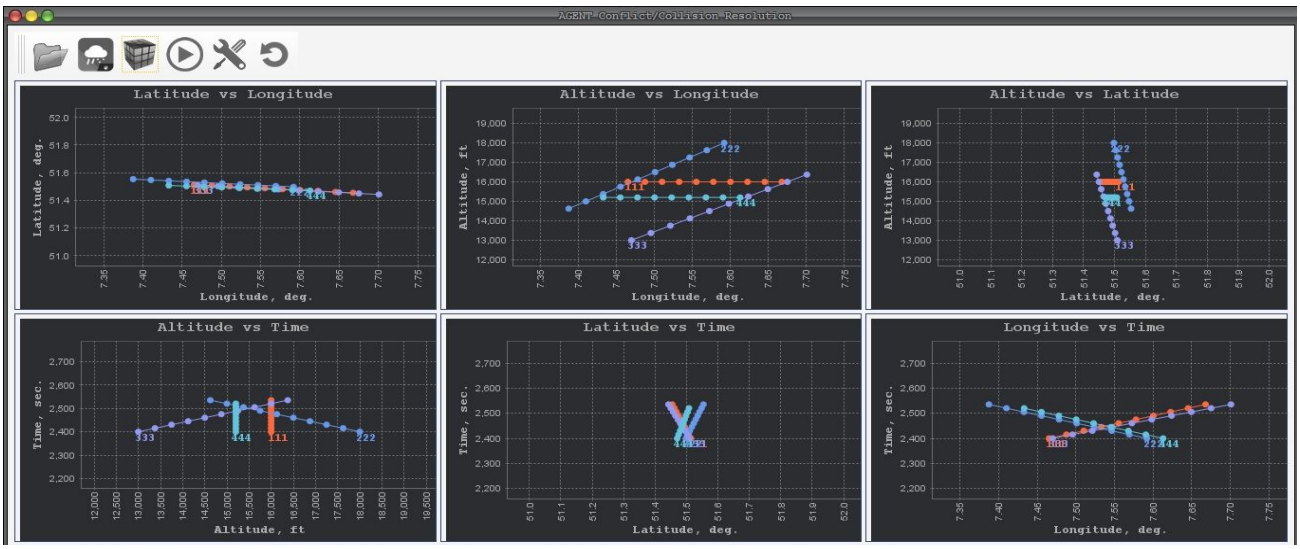


Figure 7. Stratway GUI Output.

events (Munoz *et al*, 2013) considering a larger time horizon in form of a *look-ahead* time (LAT), which is decomposed into a sequence of shorter time intervals for the control actions. For the simulated induced collision scenario, it is proposed the LAT of 300 seconds. Since, the simulation case assumes the ideal case – the constant closure rates and timestamps - the CPA will occur after 150 seconds. In real cases, i.e. the flown trajectories that include the trajectory prediction states, the CPA presents a fluctuating point, so it may occur in less or more than 150 seconds. The LAT could be sequenced in different time intervals. In this scenario, an update rate of 30 seconds has been used, meaning that the system has been considered in the discrete moments: 120, 90 and 60 seconds before the CPA. The elapsed time in less than 60 seconds denotes the TCAS convergence area based on the TAU thresholds (the TA and RA activations). The model is based on the following pre-conditions:

- The LAT provides a prediction of collision event and a way on how to avoid the hotspot at all.
- A pre-decision process (a multi-trajectory selection) is given advantage over a decision process (RA maneuver) with respect to the aircraft performance (feasibility criteria);
- With a continuous decrease in distance to the CPA less number of the potential conflict-free trajectories is achievable.

The proposed model relies on one basic concept – protected volumes. They take a shape of the imaginary cones, ground-in horizontally, with the peaks presenting the starting points of the 300-seconds time horizon. The shortest distance within these cones is the LAT distance along x-axis. These protected volumes have been considered to denote the aircraft capability to fly in a limited airspace. The limitation reflects both laterally and vertically, in the following way:

- The maximum heading change in the horizontal plane is 30 degrees. For an easier model representation, it is used the term *gradient*, presenting a coefficient of a gradual increase of the horizontal divergence measured from the x-axis, with the beginning at an identified LAT WP;
- The maximum vertical gradients from the LAT WP, i.e. ROC/D are ±5000 ft/min.

With the shortest distance and specified gradients, it is possible to define a base of the cone computing the LAT distance. After this distance, both gradients form a base with two radiuses. This imaginary base takes a shape of an ellipse. The simulation model computes all the aircraft cones together with its proximity and/or intersections defining the hotspot volumes. The intersection volumes are defined by the aircraft cones that mutually intersect in some segments of their trajectories. The shape of these volumes depends on the trajectories geometry, and considerably on the four-time colors: an entrance time of first aircraft (*time-in*, t_{1i}) and its exit time (*time-out*, t_{1o}), as well as an entrance time of second aircraft (t_{2i}) and its exit time (t_{2o}).

Once a hotspot volume has been computed and projected, the simulation model searches for the collision states by applying the TCAS RA thresholds within the intersection volumes. Therefore, any aircraft flying within its cone, but outside the intersection area, is supposed to be in a CC state. This search also includes the neighboring aircraft trajectories for the induced collision cases. If a collision state is identified/predicted the proper RAs are issued, and the aircraft perform requested maneuvers inside their imaginary cones. The model records the pairwise collisions only. It is graphically described in Figure 8.

The elements of the model are structured as follows:

- **T1** – the first transition denoting the protected volumes construction with its guard function **GU1**; **T2** - second transition defining the intersection volumes with the guard function **GU2**; **T3** – third transition that checks out the number of collision events controlled by the guard function **GU3**;

- **P1** – the place expressing the vertical gradients; **P2** – the place expressing the lateral gradient (the heading change); **P3** – the place containing 300-seconds time window; **P4** – the place that stores the along-track distances; **P5** – the control place 1 assuring that input values are satisfied; **P6** – the place linking the transitions T1 and T2, and marking the protected volumes state; **P7** – place denoting the time matrix values; **P8** – the control place 2 checking that extracted time values are satisfied; **P9** – the place that depicts the intersection volumes; **P10** – the place containing the 4D trajectory data; **P11** – the place containing the RA thresholds; **P12** – the place that stores the pairwise checks within a set of aircraft; **P13** – the place storing the pairwise induced collisions.

5 Validation and Evaluation

Presented CPN model is deployed as an essential approach to the quantitative state space analysis of the events in which the potential conflicts can likely result in collisions. At present, the causal model has been validated with some stakeholders by means of:

- *Model Purposiveness*: the conceptual model has been validated by means of a unit test (mainly though the extreme scenarios), and all detected bugs have been removed. As a result of the meeting with the experts, some modifications to the conceptual model has been added to extend the simulation/tests targets.
- *Model Plausibility*: the level of plausibility, or the expert opinion, basically referes to two features of the model. The first considers a question of whether the model *looks logical*. This answer on this question

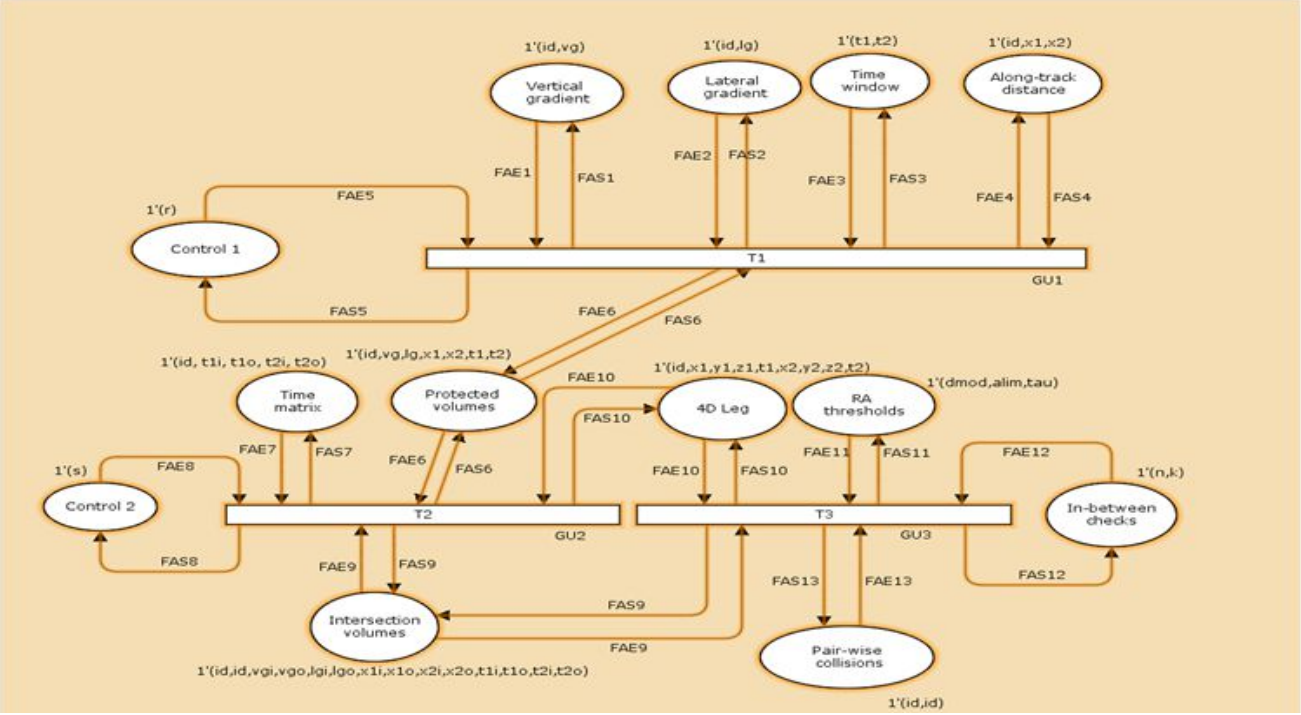


Figure 8. CPN model for pairwise collision prediction.

gives the characteristics of the model structure (the rules and hypothesis) and its parameters. The second is related to the question of whether the model *behaves logically*. This part provides an assessment of the reaction of the model outputs to the typical events (scenarios) on the inputs.

The aircraft state information, such as position and velocity, coming from the analyzed non-vectored scenario have been fed the Stratway simulation tool, and the obtained outputs – conflict segments in a form of the 4D points – have been used as an intial marking, or zero conditions, for generaion and execution of the CPN model. In addition, defined metrics, such as vertical and lateral gradients, conflict time intervals and constants (RA values) had provided a better insight of the spatiotemporal interdeoendencies in the potential collsion scenario, and creation of the intersection volumes as a qualitative solution. Finally, several simulation runs, performed in Stratway, had provided different initial markings, that are further used for computation of the final solution state in the CPN model. The intial markings pointed out to the different geometries of the conflict segments.

The follow-up validation steps will consider the state space analysis of a conflict scenario for detection of the sequence of maneuvres, that could lead to an induced collision. The generated data will be fed to InCAS (the simulation tool developed by EUROCONTROL) to validate the trajectories computed by the causal model. It will be also used *TimSpat* (Baruwa *et al*, 2015) to perform the computation of all states that can be reached from initial configuration, as illustrated in Figure 9.

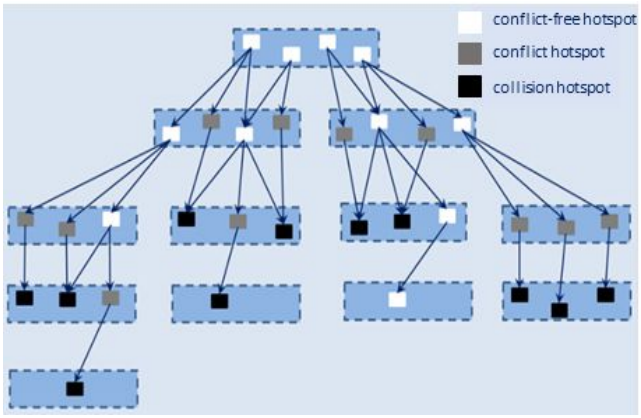


Figure 9. Reachability graph for collision events.

Each node in the graph represents a feasible marking and each arc the transition or event which allows the system evolution from the initial state to a new one. The reachability graph is structured in four levels and composed of nodes of the hotspot areas that are classified in three categories: conflict-free (white-coloured), conflict (grey-coloured) and collision (black-coloured) hotspots. There is only one node in the final level obtaining one collision hotspot. Still, the third level reaches also a node with a conflict-free hotspot.

6 Conclusions

This paper analyses the induced collision scenario in the en-route airspace as a product of the previously resolved pairwise conflicts. Based on the TCAS shortages, it tries

to identify the dynamic structures of the 4D trajectories involved in collision through simulation of their tracks and implementation of the appropriate feasible strategies. The paper further focuses on causal modeling trying to generate a new approach that will provide a higher awareness of the collision hotspot and a better decision-making process.

Acknowledgements

Research is supported by the European Union's Horizon 2020 research and innovation programme, the project *Adaptive self-Governed aerial Ecosystem by Negotiated Traffic* (under Grant Agreement No. 699313). Opinions expressed in this paper reflect the authors' views only.

References

Olatunde T. Baruwa and Miquel A. Piera. A Coloured Petri Net-Based Hybrid Heuristic Search Approach to Simultaneous Scheduling of Machines and Automated Guided Vehicles. *International Journal of Production Research*, 54(16): 4773-4792, 2015. doi: 10.1080/00207543.2015.1087656.

Thomas B. Billingsley, L. P. Espindle, and Daniel J. Griffith. TCAS Multiple Threat Encounter Analysis. *Massachusetts Institute of Technology, Lincoln Laboratory, Project Report ATC-359*, 2009.

Fabrice Drogoul, Philippe Averty, and Rosa Weber. Erasmus Strategic Deconfliction to Benefit Sesar. In: *Proceedings of the 8th USA/Europe Air Traffic Management R&D Seminar,* 2009.

ICAO. Airborne Collision Avoidance System (ACAS) Manual. *Doc 9863, AN/461*, 2006.

Tang Jun, Miquel A. Piera, and Sergio Ruiz. A Causal Model to Explore the ACAS Induced Collisions. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 228(10): 1735-1748, 2014. doi: 10.1177/0954410014537242.

Tang Jun, Miquel A. Piera, and Jenaro Nosedal. Analysis of Induced Traffic Alert and Collision Avoidance System Collisions in Unsegregated Airspace Using a Colored Petri Net Model. *Simulation*, 91(3): 233-248, 2015. doi: 10.1177/0037549715570357.

Tang Jun, Miquel A. Piera, and Olatunde T. Baruwa. A Discrete-Event Modeling Approach for the Analysis of TCAS-Induced Collisions with Different Pilot Response Times. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 229(13): 2416-2428,2015. doi:10.1177/0954410015577147.

Mykel J. Kochenderfer, Jessica E. Holland, and James P. Chryssanthacopoulos. Next-Generation Airborne Collision Avoidance System. *Lincoln Laboratory Journal*, 19(1): 17-33, 2013.

Cesar Munoz, Anthony Narkawicz, and James Chamberlain. A TCAS-II Resolution Advisory Detection Algorithm. In: *Proceedings of the AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2013.

Sergio Ruiz, Miquel A. Piera, and Isabel Del Pozo. A Medium Term Conflict Detection and Resolution System for Terminal Maneuvering Area Based on Spatial Data Structures and 4D Trajectories. *Transporation Research Part C: Emerging Technologies*, 26: 396-417, 2013. doi: 10.1016/j.trc.2012.10.005.

Multi-Sourcing and Quantity Allocation under Transportation Policies

Aicha Aguezzoul

Lorraine University, UFR ESM-IAE, Metz, France, aicha.aguezzoul@univ-lorraine.fr

Abstract

Multi-sourcing, inventory, and transportation management are among the major levers of a supply chain. In this paper, we study the case of multi-sourcing problem by considering the transportation policies used between suppliers and a buyer. Thus, we propose a programming model that determines the optimal order quantities to allocate to the suppliers used, according to the direct or indirect shipment. The objective to minimize in the model is the total logistics cost which is composed of purchasing, inventory, and transportation costs. The constraints related to suppliers and buyer are considered. The model is illustrated with a comprehensive example.

Keywords: nonlinear programming, multi-sourcing, quantity allocation, transportation, simulation

1 Introduction

Transportation and sourcing from suppliers are among the main activities for a supply chain management and optimization. Indeed, the costs of raw materials and components account for 40 to 60% of production costs for most US manufacturers (Wadhwa and Ravindran, 2007), while the transportation cost accounts a large part of the total logistics cost. The literature is very abundant on the sourcing and supplier selection studies in terms of criteria, methods for measuring their performance, and strategies on mono versus multi-sourcing.

This paper also caters these issues, and studies the impact of transportation on the multi-sourcing process. Indeed, in that process, splitting orders across multiple suppliers will lead to smaller transportation quantities which will likely imply larger transportation cost. Moreover, transportation and inventory elements are highly interrelated and contribute mostly to the total logistics cost. Finally, transportation takes a great place in the current context of sustainable development.

The remainder of this paper is organized as follows. Section 2 presents relevant literature on the sourcing strategy. In Section 3, the mathematical model is described. Section 4 reports the results of computational experiments, based on simulations. The last section contains some concluding remarks.

2 Literature review

Sourcing strategy is one of the most critical activities in purchasing process. This strategy further includes the supplier selection problem, which is widely studied in the literature. It's a very complex process because it involves various conflicting criteria such as cost, quality, delivery, reputation, flexibility, production capacity, geographical location, relationship, etc.

Dickson (1966) was the first one to address this issue and has identified 23 different criteria by which purchasing managers have selected suppliers in various procurement environments. The problem is how to select suppliers that perform satisfactorily on the desired criteria. Based on a review of 74 papers published since 1966, Weber et al. (1991) have found that vendor selection is a multi-criteria decision making (MCDM). To solve that MCDM problem, the authors showed that the proposed approaches might be grouped into three categories, which are: linear weighting methods, mathematical programming models, and statistical/probabilistic approaches. Since that time, other methods in various research contexts such as supply chain design/management, agile manufacturing, and dynamic alliances appeared in the literature. For instance: Expert system (Zhao and Yu, 2011; Amindoust et al., 2012), multi-objective programming (Aguezzoul and Ladet, 2007; Amid et al., 2009; Jadidi et al., 2014), Data Envelopment Analysis (DEA) (Sevkli et al., 2003), Analytic Network Process (ANP) (Dargi et al., 2014), hybrid approach (Demirtas and Üstün, 2009; Haldar et al., 2012; Karsak and Dursun, 2014), etc. An analysis of these methods can be found in (Aguezzoul and Ladet, 2006; Ho et al., 2010; Chai et al., 2013; Aguezzoul, 2014).

As mentioned in (Aguezzoul, 2015), in these various approaches, the criteria relating to transportation like shipment cost, geographical location of supplier, etc. are considered in the outsourcing process only in an implicit manner. Moreover, the stocks in the entire transportation network linking buyer to suppliers is not evaluated. Finally, in the current context of sustainable development, environmental dimensions such as external transport costs that are related to air pollution, noise, congestion, accidents, etc., are rarely considered.

In this paper, we refer to the work cited in (Aguezzoul, 2015) by considering the particular case of transit via a single terminal. In addition, we study the case of several scenarios depending on whether or not the suppliers use the transit via the given terminal. The detail of the model is given in the following paragraph.

3 Model development

In this article, we consider the total logistics cost as the relevant criterion to minimize in the case of sourcing from several suppliers. Such suppliers have to establish a green supply chain. The total cost is the sum of purchasing, transportation, external shipment, and inventory costs, while capacity and quality of suppliers, and buyer's demand are formulated as constraints. The model aims to deduce the order quantity for each supplier, taking into account the transportation policies to put in place between each supplier and the buyer. The following notation and formulations to model this problem are:

n: number of suppliers,

D: buyer demand per unit time, assumed constant,

Q: ordered quantity to all suppliers in each period,

Q_i: ordered quantity to i^{th} supplier in each period,

A_i: ordering cost of i^{th} supplier per order, in each period,

P_i: unit purchase price of i^{th} supplier,

C_i: production capacity of i^{th} supplier,

q_i: rate of quality of i^{th} supplier

q_a: minimum rate of quality accepted by the buyer

r: inventory holding cost rate

d_i: distance between the i^{th} supplier and the buyer,

Cf_i: shipping cost per distance between the i^{th} supplier and the buyer,

Cv_i: shipping cost per load between the i^{th} supplier and the buyer.

Cex_i: external shipping cost between the i^{th} supplier and the buyer.

The decision variables for the model are:

- X_i: fraction of Q assigned to i^{th} supplier
- $Y_i = \begin{cases} 1 \; if \; X_i > 0 \; (i^{th} \; supplier \; is \; used) \\ 0 \; if \; X_i = 0 \end{cases}$

In addition, D/Q is the number of periods during the time considered. The total cost has the following form:

$$TC = \sum_{i=1}^{n} [DX_i P_i] + \left[D/Q \begin{pmatrix} d_i Y_i (Cf_i + Cex_i) \\ +QX_i Cv_i \end{pmatrix} \right] + [rP_i Q X_i^2] \quad (1)$$

- The first term in this function is the total purchasing cost. $X_i D$ is the part of demand to assign to i^{th} supplier,
- The second term represents the total transportation cost. Cf is a fixed shipping cost which is independent of a load and includes cost of stop and cost per unit distance. Cv is a cost per load and it's independent of the distance covered. Cex is the external cost of transportation,
- The last term in the function is the total inventory cost. In a transportation network, total inventory includes loads that are waiting to be shipped from each supplier, and loads that are waiting to be used by the buyer. That supposes that each supplier produce items at a constant rate and the production planning is synchronized with that of transportation. The average time required to i^{th} supplier to produce a shipment of size Q_i is Q_i/D. Each item in the load waits on average half of this time before being shipped $Q_i/2D$. After arriving, each item waits on average $Q_i/2D$ before being used. Thus, the average time spend by an item from i^{th} supplier to buyer is: Q_i/D.

Here, we use the Economic Order Quantity (EOQ) model which is widely used to calculate the optimal lot size to reduce the total logistics cost. EOQ is the value that cancels the derivation of TC:

$$EOQ = \sqrt{\frac{D \sum_{i=1}^{n} Y_i \, d_i (Cf_i + Cex_i)}{r \sum_{i=1}^{n} P_i X_i^2}} \quad (2)$$

Thus, by replacing Q by EOQ in (1), the final expression of the total cost is:

$$TC = \sum_{i=1}^{n} DX_i (P_i + Cv_i) + 2 \sqrt{Dr \left(\sum_{i=1}^{n} d_i Y_i (Cf_i + Cex_i) \right) \left(\sum_{i=1}^{n} P_i X_i^2 \right)} \quad (3)$$

The mathematical formulation of the nonlinear programming model is given as follow:

$$Min(TC) = \sum_{i=1}^{n} DX_i (P_i + Cv_i) + 2 \sqrt{Dr \left(\sum_{i=1}^{n} d_i Y_i (Cf_i + Cex_i) \right) \left(\sum_{i=1}^{n} P_i X_i^2 \right)} \quad (4)$$

The mathematical formulation of the nonlinear programming model is given as follow:

$$Min(TC) = \sum_{i=1}^{n} DX_i(P_i + Cv_i) + 2\sqrt{Dr\left(\sum_{i=1}^{n} d_i Y_i(Cf_i + Cex_i)\right)\left(\sum_{i=1}^{n} P_i X_i^2\right)} \quad (5)$$

S.T.

$$\begin{cases} X_i D \le C_i & i = 1,n \quad (6) \\ \sum_{i=1}^{n} q_i X_i \ge q_a \quad (7) \\ \sum_{i=1}^{n} X_i = 1 \quad (8) \\ \epsilon Y_i \le X_i \le 1 & i = 1,n \quad (9) \\ Y_i = 0,1 & i = 1,n \quad (10) \end{cases}$$

- Equation (5) represents the total cost TC to minimize and whose expression is given by (4).
- Constraint (6) represents the capacity restriction for each supplier.
- Constraint (7) is an aggregate performance measure for quality for all suppliers.
- Constraint (8) indicates that demand is placed with the set of n suppliers.
- Constraint (9) requires that an order be placed with a supplier if only it's used; ε is a positive number, slightly greater than zero.
- Constraint (10) imposes binary requirements on the decision variables Y_i.

4 Numerical example

4.1 Problem Data

In this section, we present a case study of three suppliers, denoted S1, S2 and S3, who have capacities limited. Two types of shipment are used: a TruckLoad (TL) and a Less than TruckLoad (LTL), characterized respectively by the shipping cost per load of 0€ and 0.05€, the shipping cost per distance of 1.32€/mile and 0.15€/mile, and the external cost per distance of 1.11€ and 1.07€. The demand of the buyer is 1000 per week, r=20%, and the minimum accepted rate of quality is 95%. Table 1 below contains other information on the suppliers. In these experiments, we take ε = 0.001 by supposing that 0.1% is the minimum percentage of the demand that the buyer will order to a supplier.

Table 1. Suppliers' information.

Suppliers	*S1*	*S2*	*S3*
Capacity	700	800	600
Rate of quality (%)	97	95	93
Unit purchase price (€)	9	7	5
Distance to buyer (miles)	100	150	200
Distance from suppliers to terminal (miles)	50	70	100

To solve the model, simulations are used. Each one corresponds to a scenario, which depends on the mode of transport (Tl or LTL) used between each supplier and the buyer. We then have the following eight scenarios:

- Scenario 1: Each supplier uses a TL.
- Scenario 2: S1 uses a LTL while each of S2 and S3 uses a TL.
- Scenario 3: S2 uses a LTL while each of S1 and S3 uses a TL.
- Scenario 4: S3 uses a LTL while each of S1 and S2 uses a TL.
- Scenario 5: each of S1 and S2 uses a LTL while S3 uses a TL.
- Scenario 6: each of S1 and S3 uses a LTL while S2 uses a TL.
- Scenario 7: each of S2 and S3 uses a LTL while S1 uses a TL.
- Scenario 8: Each supplier uses a LTL.

4.2 Computational Results

The results presented in table 2 bellow are generated on a personal Acer computer (Intel Core, 2.10 GHz) using Matlab software version 6.1.

Table 2. Computational results.

Scenario	*% of order quantities*			*Total cost (€)*
	X_1	X_2	X_3	
1	28	37	35	7964,14
2	30	34	36	7895,71
3	22	56	22	8088,12
4	28	42	30	8320,93
5	20	55	25	8081,67
6	29	37	34	7847,01
7	23	58	19	8169,99
8	26	54	20	8030,18

From these results, we can deduce the following remarks:

- This table gives the quantity allocation for the suppliers, and the total logistics cost, according to the transportation policy considered.

- The quantities to order to each of the three suppliers depend much of the scenario used, and therefore of the considered transport policy.
- Over 50% of the quantity is allocated to the supplier within scenarios 3, 5, 7 and 8. In these cases, transportation between S2 and buyer is done by a LTL.
- In all scenarios, the total cost is minimum (= 7847, 01) for scenario 6. In this case, each of S1 and S3 uses a LTL while S2 uses a TL. The order quantities in % for suppliers are respectively 29, 37, and 34.

5 Conclusions

Multi-sourcing strategy is one of the most critical activities of purchasing management in a supply chain. A review of literature on that field shows that there has been very little work that comprehensively examines the role of the transportation in this decision.

In this paper, a nonlinear programming approach is developed to determine the order quantities to allocate to the suppliers considered, taking into account the type of transportation, LTL or TL. To solve the model, simulations are used; each one corresponds to a given scenario related to a transportation policy used to carry the products bought from the suppliers to the buyer with an aim of minimizing total logistics cost. Thus, the buyer will have available several scenarios to make its procurement decisions.

One of the prospects for this work is to integrate the case where goods can be transhipped through intermodal freight transportation. In this case, the goods undergo at least three changes of the means of transportation between their origin and their destination. For example, in a road-rail combined transport, a shipment is first transported by truck to a given terminal. Then, it is transhipped from truck to train. At the other end of the transport chain, the shipment is transhipped from train to truck and delivered by truck to the receiver.

References

A. Aguezzoul and P. Ladet. Sélection et évaluation des fournisseurs: Critères et méthodes. *Revue Française de Gestion Industrielle*, 25(2): 5-27, 2006.

A. Aguezzoul and P. Ladet. A nonlinear multiobjective approach for the supplier selection, integrating transportation policies. *Journal of Modelling in Management*, 2(2): 157-169, 2007.

A. Aguezzoul. Third-party logistics selection problem: A literature review on criteria and methods. *Omega: The International Journal of Management Science*, 49: 69-78, 2014.

A. Aguezzoul. Sourcing and transportation strategies in a supply chain: a nonlinear programming model. *In Proceedings of IEEE Service Operations and Logistics, and Informatics - SOLI'15, Tunisia, November 15-17, 2015*, pages: 160-164. doi: 10.1109/ SOLI.2015.7367612.

A. Amid, S.H. Ghodsypour and C. O'Brien. A weighted additive fuzzy multiobjective model for the supplier selection problem under price breaks in a supply chain. *International Journal of Production Economics*, 121(2): 323-332, 2009.

A. Amindoust, S. Ahmed, A. Saghafinia and A. Bahreininejad. Sustainable supplier selection: A ranking model based on fuzzy inference system. *Applied Soft Computing*, 12(6): 1668-1677, 2012.

J. Chai, J.N.K. Liu and E.W.T Ngai. Application of decision-making techniques in supplier selection: A systematic review of literature. *Expert Systems with Applications*, 40(10): 3872-3885, 2013.

A. Dargi, A. Anjomshoae, M. R. Galankashi, A. Memari and M. Binti. Supplier selection: A fuzzy-ANP approach. *Procedia Computer Science*, 31: 691-700, 2014.

E.A. Demirtas, and O. Üstün. An integrated multiobjective decision making process for supplier selection and order allocation. *Omega: The International Journal of Management Science*, 36(1): 76-90, 2008.

G.W. Dickson. An analysis of vendor selection systems and decisions. *Journal of Purchasing*, 2(1):5-17, 1966.

A. Haldar, A. Ray, D. Banerjee and S. Ghosh. A hybrid MCDM model for resilient supplier selection. *International Journal of Management Science & Engineering Managemen*t, 7(4): 284-292, 2012.

W. Ho, X. Xu and K.D. Prasanta. Multi-criteria decision making approaches for supplier evaluation and selection: A literature review. *European Journal of Operational Research*, 212(1): 16-24, 2010.

E.E. Karsak and M. Dursun. An integrated supplier selection methodology incorporating QFD and DEA with imprecise data. *Expert Systems with Applications*, 41(16): 6995-7004, 2014.

O. Jadidi, S. Zolfaghari and S. Cavalieri. A new normalized goal programming model for multi-objective problems: A case of supplier selection and order allocation. *International Journal of Production Economics*, 148: 158-165, 2014.

M. Sevkli, S.C. Lenny Koh, S. Zaim, M. Demirbag and E. Tatoglu. An application of data envelopment analytic hierarchy process for supplier selection: a case study of BEKO in Turkey. *International Journal of Production Research*, 45(9): 1973-2003, 2007.

V. Wadhwa and A.R. Ravindran. Vendor selection in outsourcing. *Computers and Operations Research*, 34(12): 3725-3737, 2007.

C.A. Weber, J. Current and W.C. Benton. Vendor selection criteria and methods. *European Journal of Operational Research*, 50(1):2-18, 1991.

K. Zhao and X. Yu. A case based reasoning approach on supplier selection in petroleum enterprises. *Expert Systems with Applications*, 38(6): 6839-6847, 2011.

A Variogram-Based Tool for Variable Selection in a Wastewater Treatment Effluent Prediction

Markku Ohenoja Jani Tomperi

Control Engineering, University of Oulu, Finland, forename.surname@oulu.fi

Abstract

In this study, a variogram method was utilized as a variable selection tool for finding the optimal subsets of variables for developing predictive models for the quality of wastewater treatment effluent. The quality of effluent was here assessed by biological and chemical oxygen demand and suspended solids in biologically treated wastewater. The dataset included, in addition to traditional process measurements, results of a novel optical monitoring device which was used for imaging an activated sludge process in-situ during a period of over one year. The study showed that the variogram based method has potential in fast and computationally easy variable selection. The developed models can be used for proactive monitoring and estimating the quality of effluent in several stages hours before in comparison to laboratory analysis taken from treated wastewater.

Keywords: activated sludge process, BOD, COD, cross-validation, modeling, optical monitoring, suspended solids, variogram

1 Introduction

While the amount of produced wastewater is increasing, the regulations for the quality of discharges by authorities are constantly tightening, and the operating costs are necessary to be minimized. Wastewaters are commonly treated in biological activated sludge processes (ASP), which are sensitive to external and internal disturbances, such as changing temperature, and varying quality and quantity of wastewater. Disturbances affect the bacterial balance of biomass and the optimum operating conditions, which are in a key role for a high pollution removal rate, low suspended solids in the effluent and a good settling properties of the sludge. Disturbances in the bacterial balance may have serious environmental and economic effects as they often produce dysfunctional flocculation and settling. The most common problem in ASP is filamentous bulking, which is caused when the secondary settler is unable to efficiently remove the suspended biomass from the wastewater. Recovery from the occurred disturbances is slow and the effects on process operation and purification result are long-lasting. (Tchobanoglous *et al*, 2003; Amaral, Ferreira, 2005; Mesquita *et al*, 2009)

On this account, an accurate operating of the wastewater purification process is required. The performance of a wastewater treatment process can be assessed analyzing the quality parameters of treated wastewater, such as biological and chemical oxygen demand (BOD, COD), suspended solids (SS), and sludge volume index (SVI). However, these parameters only show the poor quality of effluent when it already occurs and the corrective operations are inevitably late. Thus, there is a demand for new real-time monitoring tools and methods to be used in process control in parallel with the traditional offline analysis of wastewater samples and expert knowledge. The novel on-line optical monitoring method gives fast, objective information about the state of the wastewater treatment process, reveals some of the reasons for settling problems, and combined to a predictive model, shows the quality of effluent in advance (Koivuranta *et al*, 2015; Tomperi *et al*, 2017). In this study, a variogram method is utilized for finding the optimal subset of variables to develop predictive models for BOD, COD, and SS in biologically treated wastewater. The dataset from a period over one year included the results of the in-situ optical monitoring of an ASP, and the offline process measurements.

2 Material and methods

2.1 Wastewater Treatment Plant

The data used in this study was collected from the largest wastewater treatment plant (WWTP) in Finland, located in Helsinki. Viikinmäki WWTP processes daily 270,000 m^3 of wastewater from over 800,000 inhabitants around the Helsinki region. Part of the total flow (15%) come from industrial sources. This WWTP is a three-phased activated sludge process that utilizes the simultaneous precipitation method for phosphorus removal. Wastewater is processed in nine activated sludge process lines. In addition to mechanical, biological, and chemical treatment, a biological filter has been added to improve nitrogen removal. The unit operations of the process are intake, screening, grit, and grease removal, preliminary settling, aeration, degassing, secondary settling, biological de-nitrification filtration, and discharge (Figure 1). Screening removes the large solids from the water. Grit and grease removal separates rapidly settling, very coarse solids, as well as, greasy and oily substances that are lighter than water. In

the preliminary settling phase, easily settling material is separated from the water. The biological treatment is conducted by means of a de-nitrification-nitrification process in an aeration tank which is used to grow activated sludge. At the head of the aeration tank, there is a separate mixing area, where new wastewater entering the tank is reseeded with returned activated sludge from the secondary settling tank, and recycled sludge from the end of the aeration tank. Activated sludge, biomass which contains organic matter and nutrients, is separated from the treated wastewater by settling in the secondary settling tank and returned to the aeration tank. Part of the activated sludge is removed daily to maintain a suitable sludge age and sludge concentration in the aeration tank. After the secondary settling phase, wastewater is led to filtration based on bacterial action to enhance de-nitrification of the wastewater. (HSY, 2016)

2.2 Optical Monitoring and Image Analysis

To replace the slow, irregular, and subjective manual microscopic analysis of wastewater samples, a small-scale optical monitoring device and an image analysis method were developed (Koivuranta *et al*, 2013) and proved functional for monitoring the floc morphology reliably in-situ in full-scale municipal ASP (Koivuranta *et al*, 2015). The device consists of an imaging unit, a sample handling unit, and a control PC with an electronics unit. Wastewater samples were taken from one activated sludge line in the aeration tank, diluted, and pumped through a cuvette, which was imaged with a high-resolution charge-coupled device (CCD) camera. At normal flow, the delay between optical monitoring measurement and the output of the WWTP was about 13 hours. The optical monitoring device measured several morphological features of the flocs and filaments: in addition to the size parameters such as mean equivalent diameter, floc area, and filament length, the calculated shape parameters included, for example, fractal dimension, form factor, and roundness. The parameters were calculated as an average of the values for individual objects over a single image. The detailed description of the device and mathematical formulas of the calculated parameters are presented in (Koivuranta *et al*, 2013).

2.3 Variable Selection Using Variogram

Modern plants produce large amounts of data which often include irrelevant variables for a specific purpose, for instance modeling. Only significant input variables should be selected for model development. The greater number of variables does not necessary mean better prediction results because correlated, noisy and uninformative input variables increase the computational complexity, make the training of the model more difficult and worsen the prediction result. Over-fitting may occur if the model contains too many variables which are fitted not only to the data but also to the random noise. Additionally, the sampling rates of different input variables may differ significantly, therefore describing the process dynamics in different precisions.

In this work, a variogram-based method is utilized as a variable selection method in order to find the optimal subsets for modeling the suspended solids content, BOD, and COD in biologically treated wastewater. The idea of utilizing variogram for variable selection comes from the fact that a variogram of particular measurement holds the information about the relative error levels of the sampling and analysis of that measurement. Variogram is a fundamental tool within Theory of Sampling (Gy, 2004) and has already been considered in drift estimation (Paakkunainen *et al*, 2007), temporal uncertainty propagation (Jalbert *et al*, 2011), fault diagnosis (Kouadri *et al*, 2012), statistical process control (Minnit, Pitard, 2008), and as a process stability measure (Bisgaard, Kulachi, 2005).

Variogram is calculated from a set of systemically collected data. In this work, it is assumed that the data is systemically sampled and that the flow rate, or sample weight, is constant. Hence, the heterogeneity of the data can be interpreted as:

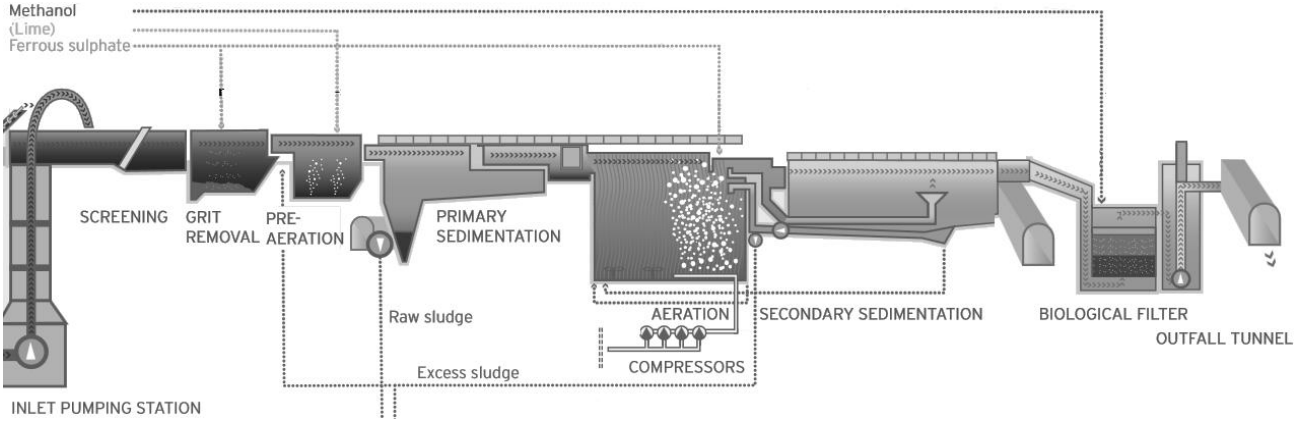


Figure 1. The wastewater treatment process at Viikinmäki. Modified from (HSY, 2016).

$$h_i = \frac{x_i - \bar{x}}{\bar{x}} \quad (1)$$

where h_i and x_i are the heterogeneity and the measurement result for sample i, respectively and $\bar{x}$ is the average of measurements x_i. The semi-variogram is calculated as:

$$v(j) = \frac{1}{2(N - j)} \sum_{i=1}^{N/2} (h_{i+j} - h_j)^2 \quad (2)$$

where $v(j)$ is the relative standard error between samples collected with lag j and N is the number of samples in the data set. The intercept $v(0)$ is estimated based on a linear extrapolation of the first $N/10$ (floored) points of the variogram. The index describing the relative information content of the measurement and thus the criterion for variable selection is calculated as relation between the estimated sampling error $v(1)$ and process variability s_P:

$$I = \frac{v(1)}{s_P} \quad (3)$$

A low value of the index I indicates that a single sample of that measurement can describe the present process variability with good accuracy. On the other hand, high value for I indicate that the relative information content of the measurement is low either due to higher sampling error or lower variability in the process.

2.4 Modeling

A k-fold cross-validation is a typical resampling method for predicting the fit of a model for a validation set, when dataset is small, and the split to separate training and validation subsets is not possible without a significant loss of data. Efficient training and validation require long and representative subset of data for both. In environmental related processes, the source dataset for model training should also encompass at least one full year of measured data because the temperature and rainfall, for instance, change depending on the season of the year and affect the process. In k-fold cross-validation, the original dataset is randomly partitioned into k subsets of equal size. One subset is used as a validation data for testing the model and the remaining k–1 subsamples are used as training data. The cross-validation process is repeated k times and each of the subsets is used only once as the validation data. A single estimation is then produced by combining these k results of the folds. Optimal k is often reported being between five and ten folds because statistical performance does not increase notably for larger values of k, and averaging over less than ten splits is computationally feasible. In this study, five-fold cross-validation was used. Multivariable linear regression (MLR) was used to predict an output variable as a linear combination of selected input variables as:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_n X_n + e \quad (4)$$

where b_o is a constant value, $b_1 \ldots b_n$ are the n regression coefficients, $X_1 \ldots X_n$ independent variables and e is the error. The performance of the model was evaluated by using Root Mean Square Error (RMSE) and coefficient of determination (R^2), which can be used to compare the relative performance of the models. (Rao *et al*, 2008; Arlot, Celisse 2010)

3 Results and Discussion

The dataset used in this work consisted of optical monitoring results and wastewater treatment process measurements from a period of over one year. On-line optical monitoring measurements were carried out at least once a day, but the laboratory measurements, on the other hand, were done only two to three times a week. During the process maintenance stoppages or occasional problems with the device, the optical measurements could not be performed. The missing laboratory and on-line data was not interpolated in this study. Thus, the total number of data points was 94 observations for 50 variables. Measurement data was scaled to range [-2, 2] before variable selection as in (Tomperi *et al*, 2017). Only variables that are useful and reliable to measure were selected. The variables from as early stage of the process as possible were preferred in order to establish models which could give proactive information of the quality of biologically treated wastewater.

Table 1. Variable selection using variogram.

Variable	Value of criterion
Fractal dimension[1]	0.17
Aspect ratio[2]	0.18
Temperature[3]	0.22
Median area of objects[4]	0.25
Filament length[5]	0.27
Roundness[6]	0.30
Sludge age[7]	0.31
Amount of filaments	0.32
Suspended solids	0.33
Number of small objects	0.34

In this study, variogram was utilized as a variable selection tool for searching the optimal subset of variables for model development. The variogram-derived indices and ten first selected variables are presented in Table 1. As seen, the most of the variables are on-line optical monitoring variables and only three of ten variables are process measurements. In comparison, five other variable selection methods tested

in (Tomperi *et al*, 2017), resulted as a suspended solids model with only one on-line optical monitoring variable (fractal dimension) and six process measurements (influent total nitrogen and sulphate, mechanically treated wastewater iron and nitrate nitrogen, temperature and anoxic proportion). Plausible reason is the lower variance of the optical monitoring variables, which shows in variogram as lower error-estimate and lower value of criterion. The earlier data analysis also showed that the quality parameters of biologically treated wastewater (BOD, COD, SS) have high mutual correlation and follow the changes of the temperature: the quality of treated wastewater was good in summer time when wastewater was warmer. The optical monitoring parameters also have several mutual correlations. For example, at summer time the amount and length of filaments was low, flocs were larger, the roundness of flocs was higher and the number of objects was lower (Tomperi *et al*, 2017). Hence, several variables in Table 1 have high mutual correlations which affect the results of model validation.

Seven input variables (n=7, 1-7 in Table 1) were selected for developing linear models for suspended solids, BOD and COD in biologically treated wastewater. The fitness of the models was predicted using 5-fold cross-validation. The results of modeling, the R^2 and RMSE values, and the regression coefficients of each developed model, are presented in Table 2. The results presented here can be considered satisfactory although in (Tomperi *et al*, 2017) the R^2 values of the SS model were between 0.79 (received using the genetic algorithm subset variable selection) and 0.71 (received using the correlation based variable selection), and the RMSE values were between 0.47 and 0.55. In the same study, the R^2 values of the BOD model were between 0.55 and 0.45, and R^2 values of the COD model were between 0.56 and 0.45.

The performance of variogram-based model for suspended solids in biologically treated wastewater is presented in Figure 2, together with the correlation-based model from the earlier study (Tomperi *et al*, 2017). All models developed using input variables selected by the variogram method have the most difference to the measured value of BOD, COD and SS at the same point: between 10-20 data point and around 40 and 75 data point. The visual interpretation also indicates that the modeling results with the variogram based variable selection contains less fast fluctuations.

The results of this study show that the variogram based tool has potential in selecting input variables for developing predictive models of treated wastewater quality even though the performance of the models was not as high as in the earlier study. Expert knowledge is required to improve the performance of the models. However, it should also be noted that the computational effort of variogram-based variable selection was minimal (less than 0.5 sec.) and implementation of the method was considerably easier than for example with genetic algorithm and successive projections algorithm, whose computational time was tens of minutes. Although the variogram-based variable selection has limited performance in the tested dataset, the method is seen interesting as it could also be developed into a recursive variable selection method due to its computational performance.

Table 2. The modeling results and the regression coefficients of input variables

Variable	R^2	RMSE	b_0	b_1	b_2	b_3	b_4	b_5	b_6	b_7
BOD	0.45	0.71	-0.64	0.47	0.67	0.09	-0.08	0.60	0.12	-0.11
COD	0.37	0.75	-0.35	-0.43	0.39	0.22	0.02	0.37	0.49	-0.13
SS	0.60	0.64	-0.61	-0.47	0.33	-0.25	0.15	0.52	0.67	-0.20

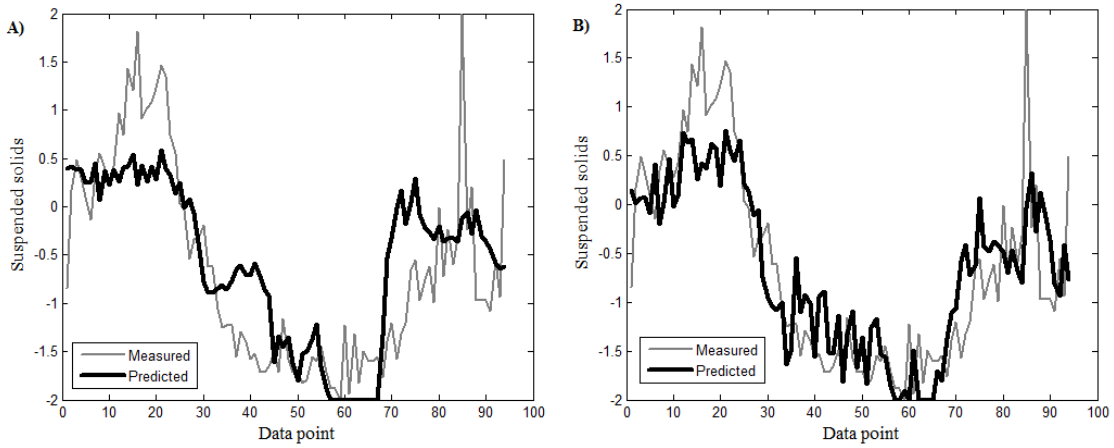


Figure 2. Measured and predicted suspended solids in biologically treated wastewater as scaled values, A) variogram based selected variables, B) correlation based selected variables (Tomperi *et al*, 2017).

4 Conclusions

In this study, a variogram method was utilized as a variable selection tool. Selected variables were used as input variables in predictive models of BOD, COD and suspended solids, which are important and critical quality parameters of the wastewater treatment process efficiency. Dataset included process measurements and the results of a novel optical monitoring method from a period of one year. Five-fold cross-validation was used to evaluate the performance of the developed models.

The presented results of variable selection show that the variogram based tool has potential in selecting input variables for developing predictive models of treated wastewater quality even though the fitness of the developed models was not as high as in the earlier study. The variogram method is, however, easier to implement and faster to use than some traditional variable selection methods. Nevertheless, the results can be considered satisfactory and the developed models can be used for proactive monitoring and estimating the quality of treated wastewater in several stages hours before in comparison to laboratory analysis taken from the treated water.

Acknowledgements

This research was carried out as part of the Measurement, Monitoring and Environmental Efficiency Assessment (MMEA), the research program of CLEEN Ltd. – Cluster for Energy and Environment.

Ms. Elisa Koivuranta (Fibre and Particle Engineering, University of Oulu) and Ms. Anna Kuokkanen (Helsinki Region Environmental Services Authority) are acknowledged for producing the original data used in this study.

References

A.L. Amaral and E.C. Ferreira. Activated sludge monitoring of a wastewater treatment plant using image analysis and partial least squares regression. *Analytica Chimica Acta*, 544:246–253, 2005. doi: 10.1016/j.aca.2004.12.061.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010. doi: 10.1214/09-SS054.

S. Bisgaard and M. Kulahci. Checking process stability with the variogram. *Quality Engineering*, 17:323–327, 2005. doi: 10.1081/QEN-200056505.

P. Gy. Sampling of discrete materials: III. Quantitative approach—sampling of one-dimensional objects. *Chemometrics and Intelligent Laboratory Systems*, 74:39–47, 2004. doi: 10.1016/j.chemolab.2004.05.011.

HSY Viikinmäki wastewater treatment plant webpage, May 2016 https://www.hsy.fi/en/experts/water-services/wastewater-treatment-plants/viikinmaki/Pages/default.aspx.

J. Jalbert, T. Mathevet. and A. Favre. Temporal uncertainty estimation of discharges from rating curves using a variographic analysis. *Journal of Hydrology*, 397:83–92, 2011. doi: 10.1016/j.jhydrol.2010.11.031.

E. Koivuranta, J. Keskitalo, A. Haapala, T. Stoor, M. Sarén and J. Niinimäki. Optical monitoring of activated sludge flocs in bulking and non-bulking conditions. *Environmental Technology*, 34:679–686, 2013. doi: 10.1080/09593330.2012.710410.

E. Koivuranta, T. Stoor, J. Hattuniemi and J. Niinimäki. On-line optical monitoring of activated sludge floc morphology. *Journal of Water Process Engineering*, 5:28–34, 2015. doi: 10.1016/j.jwpe.2014.12.009.

A. Kouadri, M. Aitouche and M. Zelmat. Variogram-based fault diagnosis in an interconnected tank system. *ISA Transactions*, 51:471–476, 2012. doi: 10.1016/j.isatra.2012.01.003.

D.P. Mesquita, O. Dias, A.L. Amaral and E.C. Ferreira. Monitoring of activated sludge settling ability through image analysis: validation on full-scale wastewater treatment plants. *Bioprocess Biosyst Eng*, 32:361–367, 2009. doi: 10.1007/s00449-008-0255-z.

R. Minnitt. and F. Pitard. Application of variography to the control of species in material process streams: %Fe in an iron ore product. *Journal of SAIMM*, 108:109–122, 2008.

M. Paakkunainen, S. Reinikainen and P. Minkkinen. Estimation of the variance of sampling of process analytical and environmental emissions measurements. *Chemometrics and Intelligent Laboratory Systems*, 88:26–34, 2007. doi: 10.1016/j.chemolab.2006.11.001.

R.B. Rao, G. Fung and R. Rosales. On the dangers of cross-validation: An experimental evaluation. *Proceedings of the 2008 SIAM International Conference on Data Mining*, Atlanta, GA, pp. 588–596, 2008. doi: 10.1137/1.9781611972788.54.

G. Tchobanoglous, F.L. Burton and H.D. Stense. *Wastewater Engineering: Treatment and Reuse*, 4th ed., Boston: McGraw-Hill, 2003.

J. Tomperi, E. Koivuranta, A. Kuokkanen and K. Leiviskä. Modelling effluent quality based on a real-time optical monitoring of the wastewater treatment process. *Environmental Technology*, 38:1–13, 2017. doi: 10.1080/09593330.2016.1181674.

Water Content Analysis of Sludge using NMR Relaxation Data and Independent Component Analysis

Mika Liukkonen[1] Ekaterina Nikolskaya[2] Jukka Selin[2] Yrjö Hiltunen[2]

[1]Department of Environmental and Biological Sciences, University of Eastern Finland, Finland,
mika.liukkonen@uef.fi

Fiber Laboratory, South-Eastern Finland University of Applied Sciences, Finland,
{ekaterina.nikolskaya,jukka.selin,yrjo.hiltunen}@xamk.fi

Abstract

In wastewater treatment, the dewatering of sludge is one of the most important steps, because it affects largely in both the process economics and the costs of sludge disposal. To optimize the dewatering processes, it would be beneficial to be aware of the different water types present in the sludge. In addition to free water, generally there are also mechanically, physically and chemically bound water within the sludge. All these water types behave differently when the sludge is dried, and they all require a different amount of energy when being removed. In this study, the Independent Component Analysis (ICA) method has been applied to an analysis of NMR (Nuclear Magnetic Resonance) relaxation data obtained from the measurement of wastewater sludge samples with a known moisture content. The results strongly suggest that the ICA method can be used for determining the amount of different water types within the wastewater sludge without a priori knowledge on their shares.

Keywords: independent component analysis, water content, nuclear magnetic resonance, sludge, relaxation decay

1 Introduction

Sludge is a semi-solid by-product remaining after wastewater treatment, industrial or refining processes. It is a separated solid suspended in a liquid, characteristically comprising large quantities of interstitial water between its solid particles (Global Water Community, 2015). This material can be dried to reduce its volume and to remove most of the moisture content of the solids within the sludge (Global Water Community, 2015). In wastewater treatment, the dewatering of sludge is one of the most important steps, because it affects largely both the process economics and the costs of sludge disposal.

It is suggested by several authors that the moisture in activated sludge can be classified to the following four categories (Kopp & Dichtl, 2000; Vesilind 1994; Tsang & Vesiling, 1990; Vesilind & Hsu, 1997; Smith & Vesiling, 1995):

- **Free** water: water which is not bound to the particles, including void water not affected by the capillary force.
- **Interstitial** water: water bound by capillary forces inside crevices and interstitial spaces of flocs.
- **Surface** water: water bound to the surface of solid particles by adhesive forces.
- **Bound** intracellular water.

This is a widely accepted classification and can be used as the reference in determining the main water types of sludge.

Another classification of water types in sludge is to divide it in three groups, i.e. 1) free water, 2) mechanically bound water, and 3) physically or chemically bound water. The free water in sludge can be easily removed by mechanical means, whereas the bound water is held firmly within the floc, bound to the sludge or trapped between the sludge particles, and thus cannot be easily removed (Jin *et al.*, 2004). The bound water can be further divided into chemically or physically bound water which is removable only by thermal drying, and mechanically bound water which is bound by weaker capillary forces (Colin & Gazbar, 1995).

In summary, it has to be emphasized that determining the water types is not straightforward, and based on the literature it is difficult to reach an unambiguous interpretation on the distribution of water within activated sludge (Vaxelaire & Cézac, 2004). Furthermore, there seem to be no studies concentrating on the analysis of water types in sludge without a priori knowledge of the shares of different water types.

Time domain nuclear magnetic resonance method (TD- NMR) is also becoming highly attractive for industrial applications due to relatively low price, mobility, easy operating, and simple sample preparation procedure. The most successful applications of TD-NMR confirmed by international standards are solid fat content determination in food and water (ISO 8292) and oil content in oilseeds (ISO 10565). They are based on the difference of NMR parameters of water and lipids and a low exchange degree between these two fractions.

A possibility to use the same principle for analysis of lipid content in microalgae (Gao *et al.*, 2008), for analysis of oil content of olive mill wastes and municipal wastewater sludge (Willson *et al.*, 2010) was demonstrated. Effects of flocculation on the bound water in sludge as measured by the NMR spectroscopy has been studied by Carberry and Prestowitz (1985).

Moreover, the international standard for hydrogen content determination in aviation fuels (ASTM D7171 – 05, 2011) has been developed recently. Metal ions, particularly paramagnetic ions, can also change significantly relaxation times in water and biological samples (Yilmaz *et al.*, 1999; Grunin *et al.*, 2013) which can be applicable when controlling wastewater treatment. Time domain NMR data have also been used in analyzing the water contents of wood and peat based fuels (Nikolskaya *et al.*, 2011) and monitoring the precipitation of metals in mine waters (Nikolskaya *et al.*, 2015).

Independent component analysis (ICA) is a statistical method that has been successfully applied to a variety of problems in signal processing (Hyvärinen *et al.*, 2001). For example, the method has been applied to a variety of problems in several fields such as brain imaging (Pulkkinen *et al.*, 2005; Calhoun *et al.*, 2002), vision research (Zhang & Mei, 2003; Ameen & Szu, 1999), telecommunications (Ristaniemi & Joutsensalo, 1999) and financial research (Kiviluoto & Oja, 1998; Back & Weigend, 1997). ICA is a method for extracting underlying, fundamental factors or components from multivariate data. It is designed so that it searches for components that are both statistically independent and non-Gaussian (Hyvärinen *et al.*, 2001), which makes it a distinguished method among the other techniques.

The complexity of spectral information can be approached by assuming that the obtained spectra are statistically independent. Principal component analysis (PCA) is the standard approach to analyze spectral data (Hyvärinen *et al.*, 2001). PCA is based on second-order statistics, which is applicable in the analysis of Gaussian distributed data. However, spectral data can comprise interesting information having a non-Gaussian distribution that can potentially be analyzed with ICA.

In the present study, the ICA method has been applied to an analysis of NMR relaxation data obtained from the measurement of wastewater sludge samples with a known moisture content.

2 Materials and methods

2.1 NMR measurements

The seven sludge samples (See Table 1) were obtained from an industrial waste water treatment plant. The samples were gathered after the dewatering stage of the process. The water contents of samples were measured using the standard oven drying method.

Relaxation times measurements were done using a mobile NMR device with a 1H resonance frequency of 25.7 MHz (Resonance Systems Ltd). The device has been modified for online measurements in industrial conditions. The permanent magnet of 0.6 T has dimension of 140x190x150 mm weighting 19 kg. The diameter of sensor hole was 10 mm. CPMG (Carr-Parcell-Meiboom-Gill) pulse sequence for spin-spin relaxation time T2 measurements was used.

Table 1. Description of sludge samples

Sample ID	Water content [%]
Sample 1	54
Sample 2	68
Sample 3	75
Sample 4	79
Sample 5	83
Sample 6	85
Sample 7	89

2.2 Independent Component Analysis

It is assumed here that there are n observed signals (i.e., types of water), WS_1, WS_2, ..., WS_n in the data, which are linear combinations of m independent components, IC_1, IC_2, ..., IC_m. The equation for IC_i can be written as:

$$WS_i = a_{i1}IC_1 + a_{i2}IC_2 + \cdots + a_{im}IC_m = \sum_{j=1}^{m} a_{ij}IC_j \quad (1)$$

where i = 1, 2, ..., n and the a_{ij} are real coefficients (contributions of ICs). The independent components, IC_j, and also the corresponding coefficients, a_{ij}, are unknown.

The statistical model in Eq. (1) is called the independent component analysis model (Hyvärinen *et al.*, 2001). The ICA model is a generative model that describes how the observed data are generated by a process of mixing the components IC_i. Both IC_i and a_{ij} need to be estimated using the observed data. The starting point for ICA is the assumption that the components IC_i are statistically independent, which can be concluded from non-gaussianity (Hyvärinen *et al.*, 2001). Here, a fixed-point algorithm (Fast-ICA) was used as an implementation of ICA (Hyvärinen *et al.*, 2001). The analysis was performed using the Fast-ICA toolbox under the Matlab software platform (Mathworks, Natick, MA, USA).

After the analysis, the relative shares of each component can be calculated using the following formula:

$$a_{ij,rel} = \frac{a_{ij}}{\sum_{j=1}^{m} a_{ij}} \times 100 \quad (2)$$

3 Results

The original NMR measurement signals can be seen in Fig. 1. It can be seen that there is a clear dependency between the amplitude of the signal and the moisture content of sludge samples. The gained NMR relaxation data were then analyzed by the ICA method based on the Hyvärinen's fixed-point algorithm (Hyvärinen *et al.*, 2001). Several numbers of ICs were tested, and three ICs were eventually used, because this setting was found to yield the most consistent and stable results. The three independent components (IC) can be seen in Fig. 2. It can be seen that all three ICs have their own, independent behavior.

According to theory, the share of bound water from the total amount of water remains stable in the sludge when the water content is increased from 0 on to a certain point (See Fig. 3, above). After this point, other types of water start to accumulate. When moisture content is 100%, all water is considered to be in a free form, but when the sludge is dried, the share of free water decreases dramatically, and the share of bound water increases. The share of the so called interstitial water (bound by weaker capillary forces) reaches its highest value at around 70 – 90 % moisture content.

In Fig. 3 (below), the calculated relative shares of different water types as a function of the total water content of the samples can be seen. It can be seen that the measured and analyzed values roughly follow the theoretic values and thus support them.

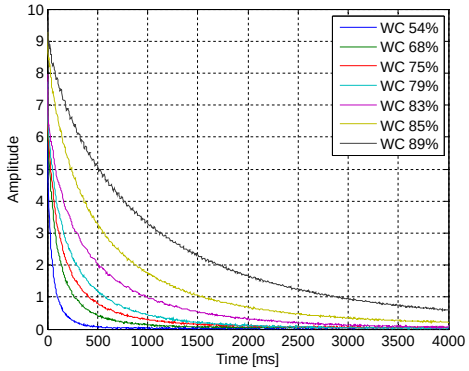


Figure 1. The observed NMR measurement signals (WS) from the 7 samples. WC = the water content of sample.

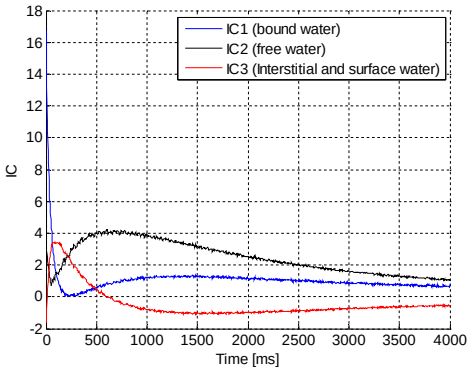


Figure 2. The three independent components (IC) computed from the NMR relaxation data.

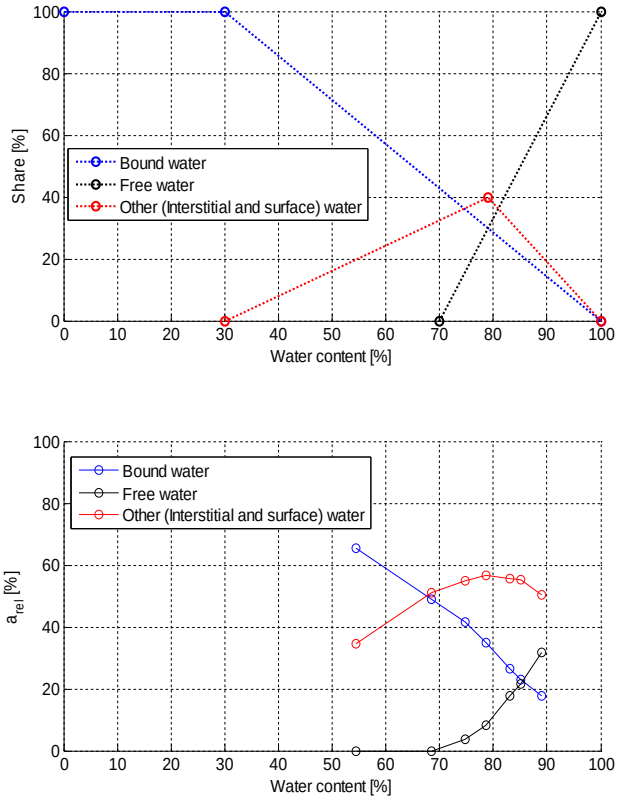


Figure 3. Theoretic (above) and calculated (below) relative shares of different water types as a function of the total water content of the samples.

4 Discussion

Measurement of the different water types in sludge is an exceptionally challenging problem, and to our knowledge this has not been tried before. In this respect, the results are extremely promising.

ICA is a universal statistical technique in which observed data are linearly transformed into components that are maximally independent from each other. A key issue in using the ICA method is to decide the number of ICs to be estimated. For the data set used, only the physically meaningful components were chosen. Our results support the use of three independent components in this case. This suggests that there are three types of signals in this data.

There is no universal truth on how many water types are present in sludge. The four water types including free, interstitial, surface and bound water, are widely accepted, but also other viewpoints exist. In this particular case three independent components could be most easily extracted from the NMR relaxation data. This suggests that there are three signals that are maximally and statistically independent when it comes to their spectra, but this does not mean that there could be more water types present as well.

Based on the results it seems that the combination of time-domain NMR and ICA can be used for determining the amount of different water types within the

wastewater sludge. It is also beneficial that the ICA method does not require a priori knowledge on the water types and their shares in the sludge. This makes it very specific and a promising approach to optimize the dewatering processes of sludge.

5 Conclusions

Based on the results it can be concluded that incorporating ICA into data analysis allows for decomposition of independent, systematically occurring patterns in NMR relaxation data. This new information can be used for guiding further study and may lead to a way of extracting the shares of different water types in wastewater sludge. This would help in making the sludge dewatering more economical and in reducing the costs of sludge disposal.

Acknowledgements

This research is a part of the *InDiGO!* (Intelligent Software and Service Concept of the Industrial Internet) project, which is funded by the Finnish Funding Agency for Technology and Innovation (*TEKES*). Furthermore, Savonlinna Smart Demonstrations (*SMD*) project (funded by the South Savo Regional Council and the European Regional Development Fund, *ERDF*) is acknowledged.

References

Global Water Community. *Sludge Drying Overview - Treatment Methods and Applications*. 2015. Available via: http://www.iwawaterwiki.org/xwiki/bin/view/Articles/SludgeDryingOverview-TreatmentMethodsandApplications

J. Kopp and N. Dichtl. The Influence of Free Water Content on Sewage Sludge Dewatering. In: *Chemical Water and Wastewater Treatment* VI, H.H. Hahn, E. Hoffmann and H. Ødegaard, Eds. Berlin Heidelberg New York: Springer-Verlag, 2000, pages 347-356.

P.A.Vesilind. The role of water in sludge dewatering. *Water Environ Res*, 66(1):4-11, 1994.

K.R. Tsang and P.A. Vesilind. Moisture distribution in sludges. *Water Sci Technol*, 22(12):135-142, 1990.

P.A. Vesilind and C.C. Hsu. Limits of sludge dewaterability. *Water Sci Technol*, 36(11):87-91, 1997.

J.K. Smith and P.A. Vesilind. Dilatometric measurement of bound water in wastewater sludge. *Water Res*, 29(12):2621-2626, 1995.

B. Jin, B.-M. Wilén and P. Lant. Impacts of morphological, physical and chemical properties of sludge flocs on dewaterability of activated sludge. *Chemical Engineering Journal*, 98:115-126, 2004.

F. Colin and S. Gazbar. Distribution of water in sludges in relation to their mechanical dewatering. *Water Res*, 29:2000-2005, 1995.

J. Vaxelaire and P. Cézac. Moisture distribution in activated sludges: a review. *Water Res*, 38:2215-2230, 2004.

ISO 8292: *2008(en) Animal and vegetable fats and oils - Determination of solid fat content by pulsed NMR*. 2008.

ISO 10565: *1998 Oilseeds - Simultaneous determination of oil and water contents - Method using pulsed nuclear magnetic resonance spectrometry*. 1998.

C. Gao, W. Xiong, Y. Zhang, W. Yuan and Q. Wu. Rapid quantitation of lipid in microalgae by time-domain nuclear magnetic resonance. *Journal of Microbiological Methods*, 75:437-440, 2008.

R.M. Willson, Z. Wiesman and A. Brenner. Analyzing alternative bio-waste feedstocks for potential biodiesel production using time domain (TD)-NMR. *Waste Management*, 30:1881-1888, 2010.

J. Bower Carberry and R.A. Prestowitz. Flocculation Effects on Bound Water in Sludges as Measured by Nuclear Magnetic Resonance Spectroscopy. *Applied and Environmental Microbiology*, 49(2):365-369, 1985.

ASTM D7171 - 05: 2011 *Standard Test Method for Hydrogen Content of Middle Distillate Petroleum Products by Low-Resolution Pulsed Nuclear Magnetic Resonance Spectroscopy*. 2011.

A. Yilmaz, M. Yurdakoc and B. Isik. Influence of transition metal ions on NMR proton T1 relaxation times of serum, blood, and red cells. *Biological Trace Element Research*, 67:187-193, 1999.

L. Grunin, E. Nikolskaya and J. Edwards. The use of 1H-NMR Relaxation Times of Water Adsorbed on Soils to Monitor Environment Pollution. *Air, Soil and Water Research*, 6:115-119, 2013.

E. Nikolskaya, M. Liukkonen, R.A. Kauppinen, L. Grunin and Y. Hiltunen. Water contents of Wood and Peat Based Fuels by Analyzing Time Domain NMR data. In: E. Dahlquist, Ed., *Proc. the 52nd International Conference of Scandinavian Simulation Society*, SIMS 2011, paper 6, 2011.

E. Nikolskaya, M. Liukkonen, J. Kankkunen and Y. Hiltunen. A non-fouling online method for monitoring precipitation of metal ions in mine waters. *IFAC Proceedings Volumes* (IFAC papers online), 48(17):98-101, 2015.

A. Hyvärinen, J. Karhunen and E. Oja. *Independent Component Analysis*. John Wiley & Sons, New York. 2001.

J. Pulkkinen, A. Häkkinen, N. Lundbom, A. Paetau, R. Kauppinen and Y. Hiltunen. Independent component analysis to proton spectroscopic imaging data of human brain tumours. *European Journal of Radiology*, 56(2):160-164, 2005.

V.D. Calhoun, T. Adali, G.D. Pearlson, P.C.M. van Zijl and J.J. Pekar. Independent component analysis of fMRI data in the complex domain. *Magnetic Resonance in Medicine*, 48:180-192, 2002.

L. Zhang and J. Mei. Shaping up simple cell's receptive field of animal vision by ICA and its application in navigation system. *Neural Networks*, 16(5-6):609-615, 2003.

M. Ameen and H. Szu. Early vision image analyses using ICA in unsupervised learning ANN. In: *Proc. International Joint Conference on Neural Networks* (IJCNN '99), vol. 2, pages 1022-1027, 1999.

T. Ristaniemi and J. Joutsensalo. On the performance of blind source separation in CDMA downlink. In: *Proc. Int. Workshop on Independent Component Analysis and Signal Separation* (ICA'99), pages 437-441, 1999.

K. Kiviluoto and E. Oja. Independent component analysis for parallel financial time series. In: *Proc. Int.Conf. on Neural Information Processing* (ICONIP'98), vol. 2, pages 895-898, 1998.

A. D. Back and A. S. Weigend. A first application of independent component analysis to extracting structure from stock returns. *Int. J. on Neural Systems*, 8(4):473-484, 1997.

Firing Accuracy Analysis of Electromagnetic Railgun Exterior Trajectory Based on Sobol's Method

Dongxing Qi, Ping Ma*, Yuchen Zhou

Control and Simulation Center, Harbin Institute of Technology, Harbin, P.R. China
QiDongxing01@163.com, PingMa@hit.edu.cn, ZhouYuchen-01@163.com

Abstract

Firing accuracy is an important index in the performance evaluation of electromagnetic railgun (EMRG). Based on a Six-DOF (degree of freedom) computer model of exterior trajectory, Sobol's method, a global sensitivity analysis approach, is utilized to analyse the influence of multiply model inputs with uncertainty on the strike accuracy of EMRG projectile. The method utilizes the firing data error and the dispersion error as the firing accuracy assessment factors of the projectile, and the input data are sampled based on Latin Hypercube Sampling (LHS). Furthermore, an example is provided, in which Sobol's method is applied in the analysis and calculation of the exterior trajectory. First-order sensitivity and total sensitivity of each factor are obtained, and then we identify the impact mechanism and interaction of different input parameters having on firing accuracy. Finally the results verify that the method is feasible and effective in the process of performance analysis of EMRG exterior trajectory.

Keywords: Sobol's method, firing accuracy, sensitivity analysis, EMRG

1 Introduction

Electromagnetic railgun (EMRG) is a typical representative of the electromagnetic emission weapons. As a new concept of weapon-system, EMRG has the advantages of quick response, hypersonic speed, and high damage efficiency for remote ground target strike mission (Fair, 2009; Ma, 2007) compared with conventional guns. And firing accuracy is an extremely important indicator to measure the operational performance of the EMRG. Nowadays, especially with the rapid development of science and technology, the accurate attack of railguns is strongly stressed by each country's military personnel (Fair, 2005). It's extremely necessary and urgent to analyze the shooting accuracy of the EMRG accurately.

Currently, the domestic and foreign research on the analysis of firing accuracy of EMRG is still in infancy. Shared research about EMRG system firing accuracy analysis is limited. Most of the studies concentrated on the analysis of guns system problems. Guo had in-depth research in detail in artillery weapons system. Reference (Guo, 2001) described the calculation method of gun's firing accuracy and the influence of various factors on the firing accuracy. But for the EMRG system, there are few complete solutions to analyze its firing accuracy precisely. Therefore, figuring out how to analyze the EMRG firing accuracy exactly is essential.

As an important tool for modeling, sensitivity analysis can let the modeler know the influence of the model parameters and inputs to the model outputs. It can be utilized effectively in modeling, model testing, and model calibration. There are a lot of classification methods about the sensitivity analysis, among which the basic idea of Sobol's global sensitivity analysis method is to study the effect of variance of the inputs parameters to the variance of outputs, using integration to describe the sum of progressive increase items factorized by system functions, then calculate ratios of the total variance to each partial variance to get the accuracy after sampling (Sobol, 2013). Compared to other methods of sensitivity analysis, Sobol's method can calculate different order sensitivity more efficiently and accurately, and it can also get the impact index, named total sensitivity which reflects the interaction of all factors. Thereby we can analyze the influence of each factor having on EMRG firing accuracy.

Sobol's method is widely applied in the field of economy, environment and climate (Yu, 2004). In this paper, we analyze the firing accuracy of electromagnetic railgun exterior trajectory by Sobol's method. Taking six factors as the research objects, such as the quality, muzzle and velocity, the first-order sensitivity and total sensitivity of each factor are calculated by the method. In the end, the influence mechanism of various factors on firing accuracy and the interaction among them is gained. The process verifies that Sobol's method is feasible in the process of accurate analysis and provides a foundation for optimizing EMRG's performance.

2 Exterior Trajectory Simulation of Electromagnetic Railgun

The model of EMRG exterior trajectory describes the process of projectiles' movement after leaving the railgun with a high speed in the atmosphere (Keshmiri,

2004; Keshmiri, 2007). It is the foundation of exterior trajectory simulation and the characteristic analysis of trajectory. The paper considers EMRG's shooting range, high altitude flight, the variation of the earth curvature and gravitational acceleration. The Six-DOF model of projectiles is built in the paper. In the model, the projectile of the EMRG is considered as a particle, and Six-DOF motion equations are established. Those are dynamic equations of projectile's centroid, dynamic equations around the centroid, kinematics equations of the centroid, kinematics equations around the centroid and the relevant initial parameters. Run the simulation program, corresponding EMRG exterior trajectory simulation results can be got.

2.1 Simulation Framework

The diagram of the simulation process is shown in Figure 1. By setting the initial projectile conditions and flight termination conditions to limit projectiles flight process, the projectile's flying state in each simulation step is determined by calculating atmospheric parameters and flight parameters of the projectile. After confirming EMRG exterior trajectory simulation process, the MATLAB programming language is utilized to achieve it.

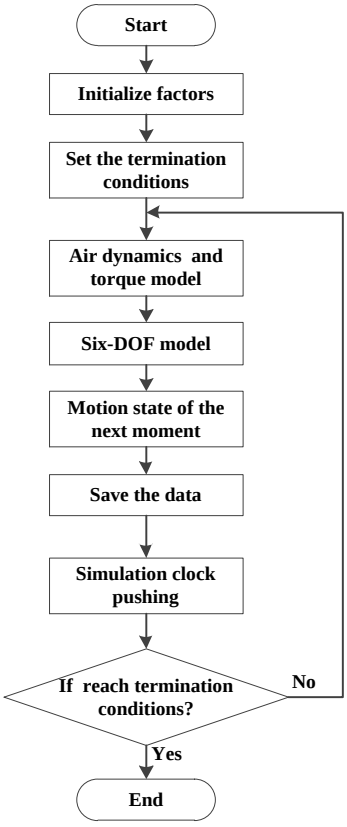


Figure 1. Electromagnetic railgun exterior trajectory simulation process.

2.2 The Experimental Design Method

In order to ensure the reliability of the projectile firing accuracy analysis which is based on the Six-DOF model, reasonable experimental design method should be chosen. Latin Hypercube Sampling (Deng, 2012) is employed in the paper. The method has great one-dimensional projection and stratified distribution characteristics, which can cover upper and lower limits of the probability distribution uniformly and distribute random number to each interval evenly. In this way, sampling frequency declines and the result remains stable. A lot of duplicate sampling work can be avoided, which improves efficiency of sampling (Ding, 2013; Zhong, 2009). The basic steps of LHS are as follows.

Set the objective function

$$y = f(x) \tag{1}$$

where y is a output variable, f is a definite function model, $x = \{x_1, x_2, \cdots, x_k\}^T$ is input variable, k is the number of input variable. Each input variable $x = \{x_1, x_2, \cdots, x_k\}^T$ subjects to a known probability distribution function $F_i(x_i)$. First of all, the input variable x_i should be sampled randomly. When random sampling, M random numbers should be generated between 0 and 1 firstly, and then transform them using the equation as follows

$$U_m = U/M + (m-1)/M \tag{2}$$

where $m = 1, 2, \cdots, M$, U is a random number between 0 and 1. U_m are the random numbers in the m'th interval.

Depending on equation(2), obviously, there is only one generated number in each interval. Because

$$(m-1)/M < U_m < m/M \tag{3}$$

where $(m-1)/M$ and m/M are the lower bound and upper bound of the m'th interval.

LHS strictly ensures the entire area was uniformly sampled, and it is almost impossible to sample repeatedly. Therefore, the method can convergence to a smaller sample size.

3 The Analysis of Electromagnetic Railgun Firing Accuracy

Electromagnetic railgun firing accuracy is influenced by multiple factors, and the influence of various factors is not identical. Meanwhile, there may be a certain coupling relationship between different factors. Thus, it is significant to analyze the impact of various factors on projectiles' firing range and direction, confirm the main factors. This section introduces the definition of firing accuracy, and gives the overall process of electromagnetic railgun firing accuracy analysis.

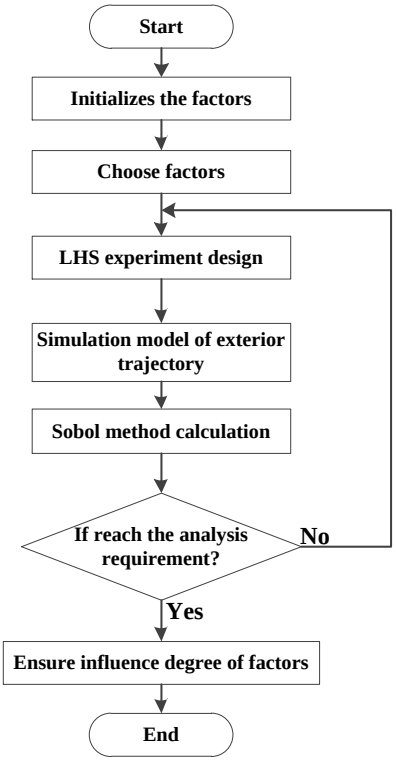


Figure 2. Process of electromagnetic railgun firing accuracy analysis.

3.1 Firing accuracy

In order to facilitate the research of the electromagnetic railgun firing accuracy, suppose that the shooting target of the electromagnetic railgun is on the ground, which is shown in Figure 3.

Assuming that the origin O is the shooting center, the deviation between projectiles average placement and shooting center is $\overline{\Delta}$, the coordinates relative to the shooting center are $(\overline{X},\overline{Z})$. The deviation between each projectile placement and shooting center is Δ, the coordinates relative to the shooting center are (x,z). The deviation of between each projectile placement and their average placement is Δ_r. For each projectile, the deviation between each projectile placement and shooting center Δ can be expressed

$$\Delta = \overline{\Delta} + \Delta_r \tag{4}$$

where Δ is the firing error, $\overline{\Delta}$ is the dispersion error, Δ_r is the firing data error. They can be written as

$$\Delta = \left[x, z\right]^T \tag{5}$$

$$\overline{\Delta} = \left[\overline{x}, \overline{z}\right]^T \tag{6}$$

$$\Delta_r = \Delta - \overline{\Delta} = \left[x_r, z_r\right]^T \tag{7}$$

where x and z are the components of the firing error in x axis and z axis; $\overline{x}$ and $\overline{z}$ are the components of the dispersion error in x axis and z axis; x_r and z_r are the components of the firing data error in x axis and z axis.

In the paper, the firing accuracy of projectiles is described by the components of the firing data error in x axis and z axis Δ_x, Δ_z and the components of the dispersion error in x axis and z axis E_x, E_z.

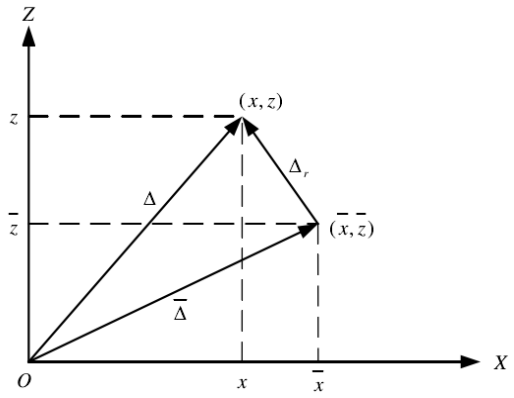


Figure 3. Projectile's firing error.

3.2 Sobol's Method

Sobol's method is based on the idea of model decomposition, 1, 2 times and higher sensitivity can be got through it (Sobol, 2013). Usually first-order sensitivity reflects the main influence parameters, and second order and higher sensitivity reflect the sensitivity among the parameters. To describe Sobol's method, firstly, one k dimension unit Ω^k should be defined as the spatial domain of input parameters (Saltelli, 2012). The main idea of Sobol's method is dividing the function $f(x)$ into some progressive increase items.

$$f(x_1,x_2,\cdots,x_k) = f_0 + \sum_{i=1}^{k} f_i(x_i) + \sum_{1\le i<j\le k} f_{i,j}(x_i,x_j) + \cdots + f_{1,2,\cdots,k}(x_1,x_2,\cdots,x_k) \tag{8}$$

where f(x) can be divided by multiple integral. In equation (8), f_0 is a constant. So we have

$$\int_0^1 f_{i_1,\cdots i_s}(x_{i_1},x_{i_2},\cdots,x_{i_s})\,\mathrm{d}\,x_{i_k} = 0 \tag{9}$$

From equation(8) and equation(9), each item is orthogonal, that is if $(i_1,i_2,\cdots,i_s) \neq (j_1,j_2,\cdots,j_l)$, then

$$\int_{\Omega^k} f_{i_1,i_2,\cdots,i_s} f_{j_1,j_2,\cdots j_l}\,dx = 0 \tag{10}$$

Thus

$$f_0 = \int_{\Omega^k} f(x)\,dx \tag{11}$$

The decomposition of equation(8) is unique, and each item can be got from multiple integral

$$f_i(x_i) = -f_0 + \int_0^1\cdots\int_0^1 f(x)\,dx{\sim}(i) \tag{12}$$

$$f_{i,j}(x_i,x_j) = -f_0 - f_i(x_i) - f_j(x_j) + \int_0^1\cdots\int_0^1 f(x)\,dx{\sim}(ij) \tag{13}$$

where $x{\sim}i$, $x{\sim}(ij)$ are the variables, except x_i and variables except x_i and x_j, respectively.

The total square deviation of f(x) is D

$$D = \int_{\Omega^k} f^2(x)dx - f_0^2 \quad (14)$$

The partial variance can be got though each items in equation (8).

$$D_{i_1,i_2,\cdots,i_s} = \int_0^1 \cdots \int_0^1 f_{i_1,i_2,\cdots,i_s}^2 \left(x_{i_1}, x_{i_2}, \cdots, x_{i_s}\right) dx_{i_1} \cdots dx_{i_s} \quad (15)$$

where $1 \le i_1 < \cdots < i_s \le k$ and $s = 1,2,\cdots,k$ .Calculate the integral of equation (15) squared in the region of Ω^k and from equation, we obtain

$$D = \sum_{i=1}^{k} D_i + \sum_{1 \le i < j \le k} D_{i,j} + \cdots + D_{1,2,\cdots,k} \quad (16)$$

Thus, sensitivity $S_{i_1,i_2,\cdots,i_s}$ can be described as

$$S_{i_1,i_2,\cdots,i_s} = D_{i_1,i_2,\cdots,i_s} / D, \left(1 \le i_1 < \cdots < i_s \le k\right) \quad (17)$$

Therefore, S_i is called the first-order sensitivity, $S_{i,j}$ is the second-order sensitivity, By that analogy total sensitivity S_{Ti} is

$$S_{Ti} = S_i + \sum_{i \ne j_1} S_{i,j_1} + \sum_{\substack{i \ne j_1 \\ i \ne j_2 \\ j_1 \ne j_2}} S_{i,j_1,j_2} + \cdots + S_{1,2,\cdots,k} \quad (18)$$

In the process of applying Sobol's method, to calculate the first-order sensitivity coefficient and total sensitivity coefficient of factors (Tarantola, 2012), assuming the gained sample $A_{N\times r}$, $B_{N\times r}$ from LHS, where N is sample size, k is the factor number, some equations are given like that

$$f_0 \approx \frac{1}{N} \sum_{j=1}^{N} f(A)_j \quad (19)$$

$$V + f_0^2 \approx \frac{1}{N} \sum_{j=1}^{N} f^2(A)_j \quad (20)$$

$$V_i + f_0^2 \approx \frac{1}{N} \sum_{j=1}^{N} f(A)_j f\left(B_A^{(i)}\right)_j \quad (21)$$

$$V_{-i} + f_0^2 \approx \frac{1}{N} \sum_{j=1}^{N} f(A)_j f\left(A_B^{(i)}\right)_j \quad (22)$$

where under the condition of A is unchangeable, change the i'th column of B into A to get $A_B^{(i)}$, and the same method to get $B_A^{(i)}$ .Thus the first-order sensitivity coefficient S_i and total sensitivity coefficient S_{Ti} of the parameter i can be written as

$$S_i = V_i / V \quad (23)$$

$$S_{Ti} = 1 - V_{-i} / V \quad (24)$$

4 Case study

In order to verify the effectiveness of the Sobol's method for the analysis of firing accuracy, in this section, the Sobol's method is applied in EMRG exterior trajectory sensitivity analysis. And the influence of each input factor on the firing accuracy is obtained. With the purpose of researching the influence of model output (ballistic range, m) according to the uncertain input factor of the model, six factors, which are mass, initial velocity, initial shooting angle, initial drift angle, y rotational angular velocities, and z rotational angular velocities may have an impact on ballistic range are extracted based on the prior knowledge, which is shown in Table 1.

Since a large number of interference parameters, in order to simulate the real environment, random wind along the positive direction of the Z axis is added to the model. After that 500 groups of testing conditions are designed by LHS. Take the series of testing conditions into the EMRG ballistic model, this series of processes are completed by simulation in MATLAB, scatter diagram is shown in Figure 1.

Table 1. Scope of Each Factor.

No.	Name of factor	symbol	units	value range
1	mass	m	kg	[9.9,10.1]
2	initial velocity	v	m/s	[1990,2010]
3	initial shooting angle	α	rad	[0.8373,0.9769]
4	initial drift angle	β	rad	[-0.001,0.001]
5	y rotational angular velocities	w_y	rad/s	[-0.001,0.001]
6	z rotational angular velocities	w_z	rad/s	[-0.001,0.001]

Table 2. Experimental Result of Longitudinal Dispersion.

Factor Names	m	v	α	β	w_y	w_z
First-order Sensitivity	0.2197	0.5851	0.5016	0.1359	0.1460	0.1440
Total Sensitivity	0.1931	0.2762	0.4432	-10^{-5}	-0.0014	0.0111

Table 3. Experimental Result of Lateral Dispersion.

Factor Names	m	v	α	β	w_y	w_z
First-order Sensitivity	-0.013	-0.012	-0.014	0.9935	-0.013	-0.014
Total Sensitivity	-0.001	0.0046	0.0013	1.0101	0.0013	10^{-5}

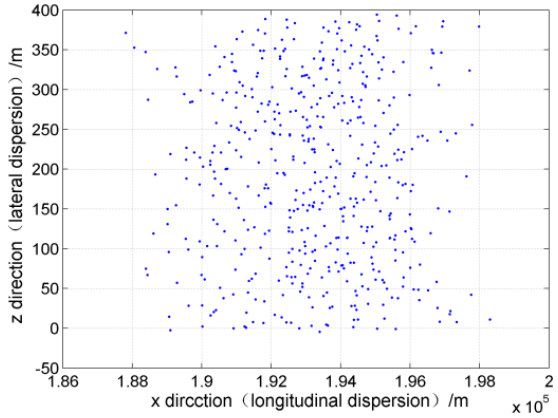


Figure 4. Scatter diagram of experiment.

The influence of each factor on the firing accuracy in two directions cannot be got from the simple scatter diagram. So we need the calculation of Sobol's method. The simulation in MATLAB needs to be run $(6\times2+1)\times500$ times. Analyze longitudinal dispersion and lateral dispersion, respectively. The experimental results are shown in and Table 3.

Transform the tables into histograms, as shown in Figures 5 and 6. From Table 2 and Figure 5, we can draw the conclusion that each factor has influence on the longitudinal dispersion and the influence degree is different. First-order sensitivity and total sensitivity of initial velocity, initial shooting angle and mass are large. Though comparing the first-order sensitivity and total sensitivity, it is explicit that influence of velocity on longitudinal dispersion decreases obviously under the comprehensive effect of multiple factors. On the contrary, the influence caused by the initial shooting angle and mass is mainly not affected by other factors. Furthermore, the sensitivity of the latter three factors is close and small. And there is poorly impact on longitudinal dispersion after they interact with other factors. In conclusion, there are three main influencing factors, initial velocity, initial shooting angle and mass under the designed experimental condition.

It is clear from Table 3 and Figure 6 that the first-order sensitivity and total sensitivity of initial drift angle all are largest (close to 1), and the sensitivity of other factors is very little (close to 0). It turns out that for the lateral dispersion, the impact of the initial drift angle is extremely significant and basically there is no effect on other factors. So in the process of testing the firing accuracy on lateral dispersion, initial drift angle should be taken into account particularly.

The first-order sensitivity in the table reflects the individual effect of each factor and the total sensitivity reflects the interaction among factors. It describes the effect of one factor on the firing accuracy under the interaction of other factors. Furthermore, the first-order sensitivity and the total sensitivity of each factor are different, which shows the interaction between the six factors is disparate.

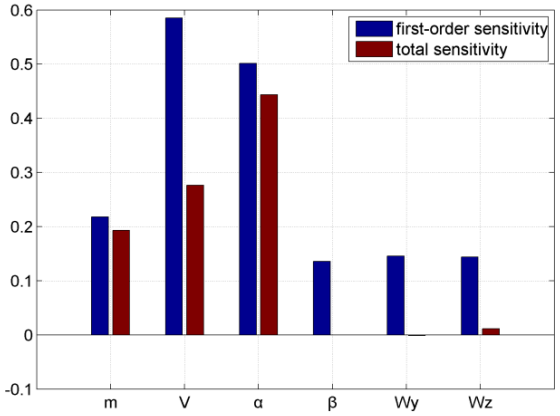


Figure 5. Result of longitudinal dispersion.

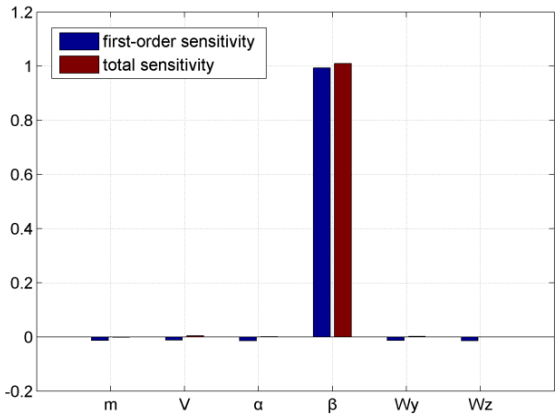


Figure 6. Result of lateral dispersion.

5 Conclusions

In this paper, the Sobol's method, a global sensitivity analysis approach, is applied to analysis of EMRG exterior trajectory firing accuracy. Single influence and interaction among each factor can be analyzed using the method, especially in the analysis which has much uncertain impact factors, a wide range of design parameters and obvious interactions between each factor. Simulation analysis results using Sobol's method show that various factors have different impact on firing accuracy of EMRG exterior trajectory, which verifies that Sobol's method is feasible and effective for firing accuracy of it. Finally the impact mechanism of each input factor on the firing accuracy and the interaction among them is obtained. In this paper, the simulation of EMRG exterior trajectory, data collection and application of Sobol's method are all implemented in MATLAB. It is easy to achieve, but the simulation needs to take a long time. In order to improve it and make firing accuracy analysis faster and more accurate, EMRG exterior trajectory simulation and data analysis tools based on C++ and Visual Studio 2010 will be designed and achieved in the future.

ACKNOWLEDGMENT

This work was supported by National Science Foundation of China (Grant No. 61627810).

References

Q. W. Deng and W. Wen. Error analysis of thin plate assembly based on Latin hypercube sampling. *China Mechanical Engineering*, 23(8):947-950, 2012. doi: 10.3969/j.issn.1004-132X.2012.08.014.

M. Ding, J. J. Wang and S. H. Li. Probabilistic load flow evaluation with extended Latin hypercube sampling. *Proceedings of the CSEE*, 33(4):163-170, 2013.

H. D. Fair. Advances in electromagnetic launch science and technology and its applications. *IEEE Transactions Magnetics*, 45(1):225–230, 2009. doi: 10.1109/elt.2008.9.

H. D. Fair. Electromagnetic launch science and technology in the United States enters a new era. *IEEE Transactions on Magnetics*, 41(1):158-164, 2005. doi: 10.1109/tmag.2004.838744.

C. Z. Fan and W. K. Wang. The development of electromagnetic railgun. *Journal of Yanshan University*, 31(5):377-386, 2007. doi: 10.3969/j.issn.1007-791X.2007.05.001.

X. F. Guo. Determination of accuracy index of firing data of fire control computer. *Journal of Ballistics*, 13(1):86-89, 2001. doi: 10.3969/j.issn.1004-499X.2001.01.018.

S. Keshmiri, R. Colgren and M. Mirmirami. Development of an aerodynamic database for a generic hypersonic air vehicle. *AIAA Guidance, Navigation, and Control conference and Exhibit*, 2005. doi: 10.2514/6.2005-6257.

S. Keshmiri and M. D. Mirmirani. Six-DOF modeling and simulation of a generic hypersonic vehicle for conceptual design studies. *AIAA Modeling and Simulation Technologies Conference and Exhibit*, 2004. doi: 10.2514/6.2004-4805.

P. Ma, Y. C. Zhou, X. B. Shang and M. Yang. Firing accuracy evaluation of electromagnetic railgun based on multicriteria optimal Latin hypercube design. *IEEE Transactions on Plasma Science*, 45(7):1503-1511, 2017. doi: 10.1109/tps.2017.2705980.

A. Saltelli, P. Annoni and I. Azzini. Variance based sensitivity analysis of model output. *Design and estimator for the total sensitivity index. Computer Physics Communications*, 181(2):259-270, 2010. doi: 10.1016/j.cpc.2009.09.018.

I. M. Sobol. Theorems and examples on high dimensional model representation. *Reliablity Engineering and System Safety*, 79(2):163-170, 2013. doi: 10.1016/s0951-8320(02)00229-6.

S. Tarantola, W. Becker and D. Zeitz. A comparison of two sampling methods for global sensitivity analysis. *Computer Physics Communications*, 183(5):1061-1072, 2012. doi: 10.1016/j.cpc.2011.12.015.

H. J. Yu and R. Li. Application of Sobol's method in sensitivity analysis of nonlinear vibration isolation system. *Journal of Vibration Engineering*, 13(2):210-213, 2004.

Z. J. Zhong and X. X. Hu. A fast precision analysis and error compensation method for missile based on Latin hypercube sampling. *Ordnance Industry Automation*, 28(6):23-25, 2009. doi: 10.3969/j.issn.1006-1576.2009.06.009.

Modelling and Simulation of a Paraglider Flight

Marcel Müller[1] Abid Ali[2] Alfred Tareilus[3]

[1]Faculty of Electrical Engineering, University of Applied Sciences Würzburg-Schweinfurt, D-97421 Schweinfurt, Germany, marcel.mueller.3@student.fhws.de
[2]Faculty of Electrical Engineering, University of Applied Sciences Würzburg-Schweinfurt, D-97421 Schweinfurt, Germany, abid.ali@fhws.de
[3]Gleitschirmfreunde Schweinfurt e.V., D-97506 Grafenrheinfeld, Germany, alfred.tareilus@web.de

Abstract

A simulation model of a paraglider flight is presented in this paper. The dynamics of the paraglider trajectory, as well as its twist angle are governed by a complex interplay between gravity and drag force. Differential equations describing this dynamical behaviour are developed by balancing all the forces and torques acting on the system. The developed model allows to compare different flight situations and find out optimal values of the parameters, which may be used in further steps for example optimization of the towing process. Simulation results of the paraglider launch process using a vehicle unreeling winch and subsequent gliding under different wind conditions are compared with real data.
Keywords: paraglider, gliding flight, towing process, aerodynamics, simulation, modelling

1 Introduction

In the 1960s, a hype about paragliding started in Europe. The performance of gliders increased significantly and the crafts were easier to control due to advanced designs. This was a major step towards making the flight safer (Currer, 2011). However, the development of modern paragliders, the ones we know today, needed a long time. There are some competitions in different disciplines as well, whereby aerial acrobatics like spiral dives or full stalls are often performed (Whittall, 2010). Most pilots drive to big mountains like the Alps to enjoy paragliding.

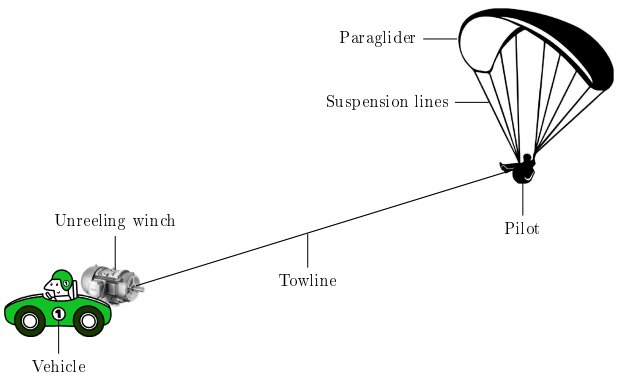


Figure 1. Winch tow launch.

Potential energy or a start altitude is transformed into kinetic energy to bridge distances as long as possible. This process is called "gliding flight". But there is an alternative method for paragliding as well. The so called winch tow launch can also be used in lowlands (see Figure 1). With the help of an unreeling winch mounted on the back of a vehicle the pilot is pulled into the air. Various parameters can be influenced during this process to reach the best possible outcome. A simulation model of the towing procedure can be helpful for optimisation of the process.

The standard reference written by Janssen et al. (2013) for paragliders and hang gliders provides guidelines about different security measures to prevent hazard situations. Consequently, it is an ideal guide for trainee pilots as well as experienced gliders. Whittall (1995) introduced basic flying techniques and presented additional information about equipment, weather and soaring. Over the years, the aerodynamics and flight mechanics of paragliders improved continuously and the fundamentals of fluid mechanics as well as the involved physical processes are nowadays well understood (Oertel, 2005). A detailed description of the effective forces acting on the paraglider and the pilot during a gliding flight can be visualised with the help of a free body diagram (Voigt, 2003). This free body diagram can be augmented in order to describe the so-called towing process (see Figure 1), whereby the pilot starts from the ground and is lifted into the air (Fahr, 1992). A polar curve, also named as Lilienthal Curve (Karbstein, 1996), describes the ratio of sink velocity to air speed in x-direction for a specific glider under different operating conditions. It allows, as a decisive advantage, to predict the glider's performance (Currer, 2011).

In spite of all the improvements and understanding of physical phenomenon and parameters, there is no complete model available, which could be employed to simulate the dynamical behaviour of the system. The aim of this contribution is to derive such a mathematical model. It could be used to simulate the system behaviour in towing and gliding states. This could lead on one hand to analyse the performance, safety and stability of paragliders and to optimise the towing process on the other hand. This paper is organised as follows. A complete mathematical model of the system is derived in section 2. How this model can be parametrised, is discussed in section 3. Sec-

tion 4 presents some simulation results, while conclusions and a brief outlook are given in section 5.

2 Modelling

In order to develop the model a simplified free body diagram of the whole system is drawn in Figure 4. The mass of the paraglider m_g and the mass of the pilot M_p are assumed to be connected to each other rigidly. The aim is to derive differential equations describing translational and rotational dynamics of the system. A coordinate system "CS" is displayed next to the figures to define the positive axes and angles.

2.1 Centre of Gravity

The centre of gravity G is the point at which the sum of all torques due to gravitational forces equals zero. The two torques involved are the torque due to the weight of the pilot and the torque due to the weight of the paraglider. Consequently, the distance l_{MG} can be determined (see equation (2)). With this information, l_{Gm} is calculable with the help of the length of the suspension lines l_{line} as well (equation (3)). The line length is known from the data sheet provided by a paraglider manufacturer which is used as a reference glider in the simulation.

$$M_p \cdot g \cdot l_{MG} = m_g \cdot g \cdot (l_{line} - l_{MG}) \tag{1}$$

$$\Leftrightarrow l_{MG} = \frac{m_g \cdot l_{line}}{M_p + m_g} = \frac{m_g \cdot l_{line}}{m_t} \tag{2}$$

$$l_{Gm} = l_{line} - l_{MG} \tag{3}$$

2.2 Aerodynamic Forces

Aerodynamics deal with forces caused by air flow around an object. Typical forces and angles of a paraglider in static conditions are shown in Figure 2. The notational symbols are explained in Table 1. The weight $F_{W,t}$ points

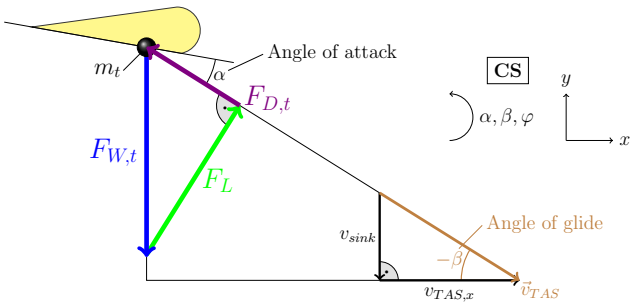


Figure 2. Aerodynamic forces.

vertically downwards and is compensated by the total aerodynamic force which is the resultant of the lift force F_L and the drag force $F_{D,t}$. They are perpendicular to each other whereby F_L creates the lift to glide through the air. The drag force $F_{D,t}$ is directed contrary to the direction of movement. Drag is caused by friction between paraglider and air and converts a part of the energy into losses. The

Table 1. Notations.

Symbol	Description
$F_{D,t}$	Drag force
F_L	Lift force
$F_{W,t}$	Weight
$T_{D,g}$	Paraglider torque caused by drag force
$T_{D,p}$	Pilot torque caused by drag force
T_L	Torque caused by lift force
T_{res}	Resultant torque clockwise
T_{start}	Start torque

half density of the air $\frac{\rho_{air}}{2}$ multiplied by the squared true air velocity $(v_{TAS})^2$ yields the dynamic pressure $p_{dynamic}$ of the air. The total drag force of a paraglider $F_{D,t}$ is equal to the dynamic pressure $p_{dynamic}$ multiplied by the projected area $A_{project}$ and the drag coefficient $c_{d,g}$ of the paraglider (Anderson, 2005).

$$F_{D,t} = c_{d,g} \cdot A_{project} \cdot \underbrace{\frac{\rho_{air}}{2} \cdot (v_{TAS})^2}_{p_{dynamic}} \tag{4}$$

Figure 2 displays how the total drag force $F_{D,t}$, under the assumption of static conditions, alternatively can be determined. It means that no acceleration force has an affect on the system. The drag force $F_{D,t}$ equals the total weight $F_{W,t}$ multiplied by the sine of the negative angle of glide $\sin(-\beta)$, which in turn is the sink velocity v_{sink} divided by the true air speed v_{TAS}.

$$F_{D,t} = F_{W,t} \cdot \sin(-\beta) = F_{W,t} \cdot \frac{v_{sink}}{v_{TAS}} \tag{5}$$

Equating equation (4) with equation (5)

$$c_{d,g} \cdot \frac{\rho_{air}}{2} \cdot A_{project} \cdot (v_{TAS})^2 = F_{W,t} \cdot \frac{v_{sink}}{v_{TAS}} \tag{6}$$

$$\Leftrightarrow c_{d,g} = \frac{F_{W,t} \cdot 2}{A_{project} \cdot \rho_{air}} \cdot \frac{v_{sink}}{(v_{TAS})^3} \tag{7}$$

The lift force F_L follows the same rule as the drag force (Anderson, 2005). The dynamic pressure $p_{dynamic}$ multiplied by the projected square $A_{project}$ and the lift coefficient c_l is equal to the lift force.

$$F_L = c_l \cdot A_{project} \cdot \underbrace{\frac{\rho_{air}}{2} \cdot (v_{TAS})^2}_{p_{dynamic}} \tag{8}$$

It is clear from Figure 2 that the lifting force F_L equals the total weight $F_{W,t}$ multiplied by the cosine of the negative angle of glide $\cos(-\beta)$ which is equal to the ratio of the horizontal velocity $v_{TAS,x}$ and the true air speed v_{TAS}.

$$F_L = F_{W,t} \cdot \cos(-\beta) = F_{W,t} \cdot \frac{v_{TAS,x}}{v_{TAS}} \tag{9}$$

Equating equation (8) with equation (9)

$$c_l \cdot \frac{\rho_{air}}{2} \cdot A_{project} \cdot (v_{TAS})^2 = F_{W,t} \cdot \frac{v_{TAS,x}}{v_{TAS}} \quad (10)$$

$$\Leftrightarrow c_l = \frac{F_{W,t} \cdot 2}{A_{project} \cdot \rho_{air}} \cdot \frac{v_{TAS,x}}{(v_{TAS})^3} \quad (11)$$

2.3 Definition of Angles

The tangent of the glide angle β, located between horizontal line and direction of movement v_{TAS} (see Figure 3), is calculated by the ratio of the air speed in y-direction to the air speed in x-direction (equation (12)).

$$\beta = \arctan\left(\frac{v_{TAS,y}}{v_{TAS,x}}\right) \quad (12)$$

The angle of attack α is the angle at which the airflow meets the wing (Currer, 2011) and is defined between the chord line and the direction of flight (see Figure 3). This angle influences the pendulum torque around the pitch axis mandatorily and can be controlled by the pilot via the brakes. The negative angle of glide $-\beta$ equals the angle

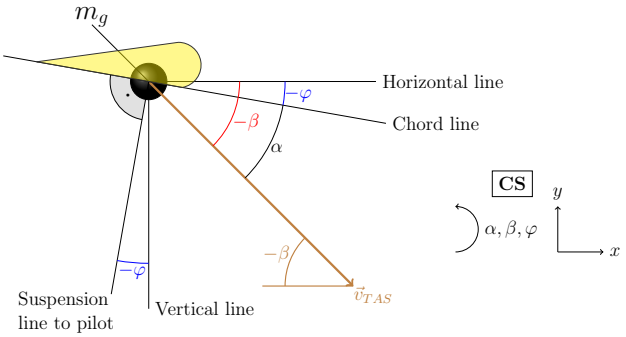


Figure 3. Definition of angles.

of attack α minus the angle of twist φ. Consequently, the angle of attack α is the difference of the angle of twist φ and the angle of glide β (see Figure 3).

$$-\beta = \alpha - \varphi \quad (13)$$

$$\alpha = \varphi - \beta \quad (14)$$

The negative angle of twist $-\varphi$ can also be found between the vertical line and the suspension line to the pilot. This information helps to determine the lever arms needed for the torque calculations (see section 2.4).

2.4 Lever Arms

As a next step the lever arms of the pendulum system have to be determined. Figure 4 clarifies the relationship between the angles of a paraglider (section 2.3) and the corresponding lever arms (section 2.4). The distances l_{Gm} and l_{MG} result from the position of the gravity point G which

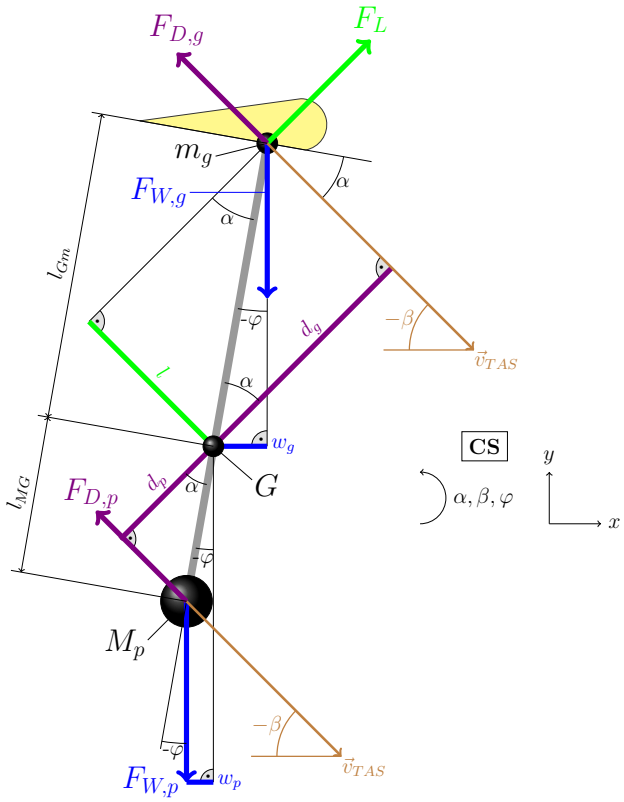


Figure 4. Lever arms.

has been calculated before in section 2.1.

$$w_g = l_{Gm} \cdot \sin(-\varphi) \quad (15)$$

$$d_g = l_{Gm} \cdot \cos(\alpha) \quad (16)$$

$$l = l_{Gm} \cdot \sin(\alpha) \quad (17)$$

$$d_p = l_{MG} \cdot \cos(\alpha) \quad (18)$$

$$w_p = l_{MG} \cdot \sin(-\varphi) \quad (19)$$

With the help of the lever arms the torques around the gravity point G can be determined (see section 2.6).

2.5 Catenary

If the pilot starts from the ground and is pulled into the air with the help of the towline, the situation slightly changes.

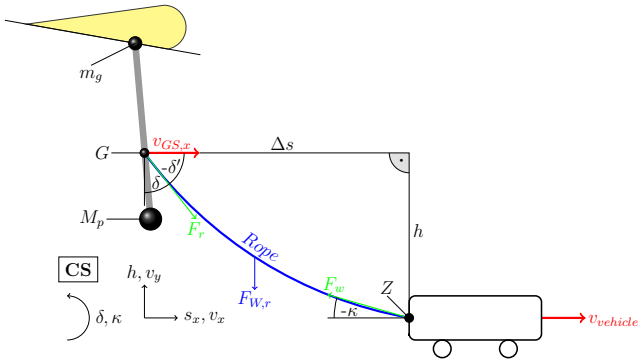


Figure 5. Catenary.

The weight $F_{W,r}$ of the rope causes a sag and results in a curved line, which directly affects the rope angles δ and κ (see Figure 5). The angles δ and κ are calculated recursively with the help of the catenary. The pilot starts from the ground and thus, the angle κ is at the beginning of the towing process zero ($\kappa_{initial} = 0°$). The catenary can be approximated by the following equation (Dankert and Dankert, 2009)

$$y(x) = C_2 \cdot x^2 + C_1 \cdot x + C_0 \tag{20}$$

where

$$C_2 = \frac{0.5 \cdot q_0}{F_w \cdot \cos(-\kappa)}. \tag{21}$$

The unknown parameters C_0 and C_1 are determined with the help of Figure 5. The height of the catenary at the position $x = 0$ equals

$$y(0\,\mathrm{m}) = C_0 = h \tag{22}$$

and at the position $x = \Delta s$

$$y(\Delta s) = C_2 \cdot \Delta s^2 + C_1 \cdot \Delta s + C_0 = 0 \tag{23}$$

$$\Leftrightarrow C_1 = \frac{-C_0}{\Delta s} - C_2 \cdot \Delta s. \tag{24}$$

Consequently the parameters C_1 and C_2 are defined. With the parameters C_0, C_1 and C_2 the rope angles δ and κ can be calculated. By differentiating equation (20) with respect to x results

$$\frac{\mathrm{d}y(x)}{\mathrm{d}x} = 2 \cdot C_2 \cdot x + C_1. \tag{25}$$

Figure 5 also displays the derivation at the position $x = 0$

$$\left.\frac{\mathrm{d}y(x)}{\mathrm{d}x}\right|_{x=0} = C_1 = \tan(\delta') \tag{26}$$

with $\delta' = \delta - \frac{\pi}{2}$ follows

$$\delta = \tan^{-1}(C_1) + \frac{\pi}{2} \tag{27}$$

as well as at the position $x = \Delta s$

$$\left.\frac{\mathrm{d}y(x)}{\mathrm{d}x}\right|_{x=\Delta s} = 2 \cdot C_2 \cdot \Delta s + C_1 = \tan(-\kappa) \tag{28}$$

$$\Leftrightarrow \kappa = -\tan(2 \cdot C_2 \cdot \Delta s + C_1). \tag{29}$$

By balancing the forces in x-direction (see Figure 5) results the rope force F_r which is needed for the following section 2.6.

$$F_r = F_w \cdot \frac{\cos(-\kappa)}{\sin(\delta)} \tag{30}$$

Until an altitude of 50 metres the constant winch force F_w is reduced to prevent large angles of attack α which could otherwise lead to dangerous flying situations. Safety is very important and gets top priority.

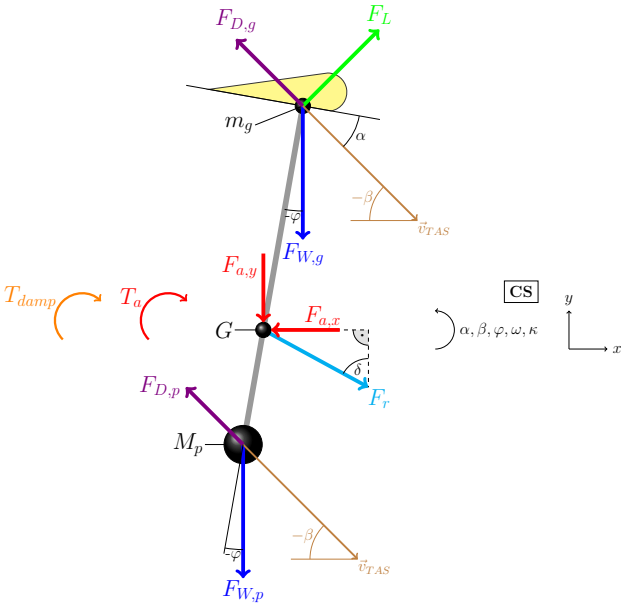


Figure 6. Balance of forces whilst towing.

2.6 Equations of Motion

Figure 6 displays the flight system whilst towing. For a gliding flight the rope force F_{rope} is zero. The algebraic sum of all forces in x-direction is zero. The acceleration force in x-direction equals the total mass m_t multiplied by the time derivative of the ground speed in x-direction $\dot{v}_{GS,x}$.

$$\begin{aligned} m_t \cdot \dot{v}_{GS,x} = &- F_{D,g} \cdot \cos(-\beta) - F_{D,p} \cdot \cos(-\beta) \\ &+ F_L \cdot \sin(-\beta) + F_r \cdot \sin(\delta) \end{aligned} \tag{31}$$

Balancing the forces in y-direction results in the following differential equation.

$$\begin{aligned} m_t \cdot \dot{v}_{GS,y} = &- F_{W,p} - F_{W,g} + F_{D,g} \cdot \sin(-\beta) \\ &+ F_{D,p} \cdot \sin(-\beta) + F_L \cdot \cos(-\beta) \\ &- F_r \cdot \cos(\delta) \end{aligned} \tag{32}$$

The algebraic sum of all torques around the gravity point G is zero. The lever arms have already been calculated in section 2.4 and are now used in equation (33). The accelerating torque around the gravity point G equals the total moment of inertia I_t multiplied by the time derivative of the angular velocity $\dot{\omega}$.

$$\begin{aligned} I_t \cdot \dot{\omega} = &+ F_{W,p} \cdot w_p - F_{W,g} \cdot w_g - F_{D,p} \cdot d_p \\ &+ F_{D,g} \cdot d_g - F_L \cdot l - T_{damp} \end{aligned} \tag{33}$$

The damping torque T_{damp} caused by the movement of the paraglider in the air equals the damping coefficient d multiplied by the angular velocity ω.

$$T_{damp} = d \cdot \omega \tag{34}$$

2.7 Velocities

The true air speed v_{TAS} is crucial to calculate the aerodynamic forces. It is the velocity of the airflow in relation to the aerofoil. In addition, the ground speed v_{GS} can be calculated with the true airspeed and the wind speed (Currer, 2011). The true air speed in x-direction $v_{TAS,x}$ is equal to the ground speed in x-direction $v_{GS,x}$ minus the headwind speed $v_{wind,x}$ (equation (35)). The true air speed in y-direction $v_{TAS,y}$ is equivalent to the ground speed in y-direction $v_{GS,y}$ minus the upwind speed $v_{wind,y}$ (equation (36)).

$$v_{TAS,x} = v_{GS,x} - v_{wind,x} \tag{35}$$
$$v_{TAS,y} = v_{GS,y} - v_{wind,y} \tag{36}$$

2.8 Self Stabilisation

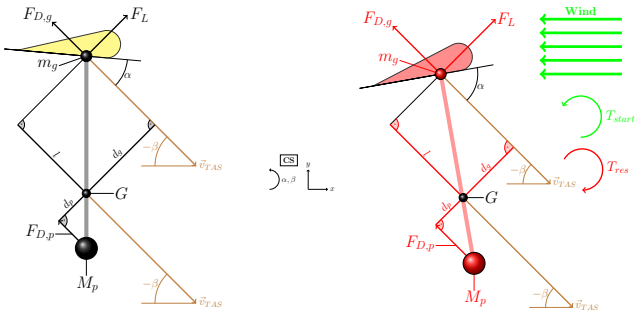


Figure 7. Self stabilisation.

A paraglider is sometimes exposed to turbulence conditions during a flight for example a sudden appearance of contrary wind (see Figure 7). The wind situation at higher altitudes is often different from the situation at ground. It is important to mention in this context that the torques due to weights cancel each other out if the paragliding system starts to swing (see equation (1)). Consequently, the weights $F_{W,g}$ and $F_{W,p}$ are not drawn in Figure 7. It displays a "standard" flight situation on the left and the reaction of the system due to headwind on the right side. This wind causes a counterclockwise torque T_{start}, which leads to an increase the angle of attack α. This affects lift and drag coefficients which are getting larger (see Figure 8). As a result, F_L and $F_{D,g}$ are also getting larger and the lever arms d_g and d_p shorter (see section 2.4). The only increased lever arm is l, which leads to an increase of T_L. This causes a self-aligning moment T_{res} trying to compensate the start torque. Consequently, the paraglider stabilises on its own and the pilot often does not have to intervene.

2.9 Complete Model

Now we summarize our discussion of previous sections and write the complete model of the system. The dynamics of six state variables of the system can be described by the following state equations.

$$\dot{s}_x = v_x \tag{37}$$
$$\dot{v}_x = \frac{1}{m_t} \cdot [-F_{D,g} \cdot \cos(-\beta) - F_{D,p} \cdot \cos(-\beta) + F_L \cdot \sin(-\beta) + F_r \cdot \sin(\delta)] \tag{38}$$
$$\dot{s}_y = v_y \tag{39}$$
$$\dot{v}_y = \frac{1}{m_t} \cdot [-F_{W,p} - F_{W,g} + F_{D,g} \cdot \sin(-\beta) + F_{D,p} \cdot \sin(-\beta) + F_L \cdot \cos(-\beta) - F_r \cdot \cos(\delta)] \tag{40}$$
$$\dot{\varphi} = \omega \tag{41}$$
$$\dot{\omega} = \frac{1}{I_t} \cdot [+F_{W,p} \cdot w_p - F_{W,g} \cdot w_g - F_{D,p} \cdot d_p + F_{D,g} \cdot d_g - F_L \cdot l - d \cdot \omega] \tag{42}$$

With

$$F_L = c_l \cdot A_{project} \cdot \frac{\rho_{air}}{2} \cdot (v_{TAS})^2 \tag{43}$$
$$F_{D,g} = c_{d,g} \cdot A_{project} \cdot \frac{\rho_{air}}{2} \cdot (v_{TAS})^2 \tag{44}$$
$$F_{D,p} = c_{d,p} \cdot A_{pilot} \cdot \frac{\rho_{air}}{2} \cdot (v_{TAS})^2 \tag{45}$$
$$F_{W,g} = m_g \cdot g \text{ and } F_{W,p} = M_p \cdot g \tag{46}$$
$$F_{D,p} = c_{d,p} \cdot A_{project} \cdot \frac{\rho_{air}}{2} \cdot (v_{TAS})^2 \tag{47}$$
$$v_{TAS,x} = v_{GS,x} - v_{wind,x} \tag{48}$$
$$v_{TAS,y} = v_{GS,y} - v_{wind,y} \tag{49}$$
$$\beta = \arctan\left(\frac{v_{TAS,y}}{v_{TAS,x}}\right) \tag{50}$$
$$\alpha = \varphi - \beta \tag{51}$$
$$\delta = \tan^{-1}(C_1) + \frac{\pi}{2} \tag{52}$$

3 Model Parameters

In this section we want to summarize important parameters used in this model. Values of some of the parameters are provided by the manufacturer or calculated analytically. Other parameters are tuned experimentally, for instant the damping coefficient d, to get a good image of the reality.

$\rho_{air} = 1.27\,\frac{\text{kg}}{\text{m}^3}$	$m_g = 6\,\text{kg}$
$M_p = 100\,\text{kg}$	$l_{line} = 6.97\,\text{m}$
$A_{projected} = 24.26\,\text{m}^2$	$A_{pilot} = 1.0\,\text{m}^2$
$l_{MG} = 0.395\,\text{m}$	$l_{Gm} = 6.575\,\text{m}$
$d = 10000\,\frac{\text{Nms}}{\text{rad}}$	$c_{d,p} = 0.33$

The aerodynamic coefficients c_l and $c_{d,g}$ are assumed to be functions of the angle of attack α. These dependencies are shown in Figure 8 graphically.

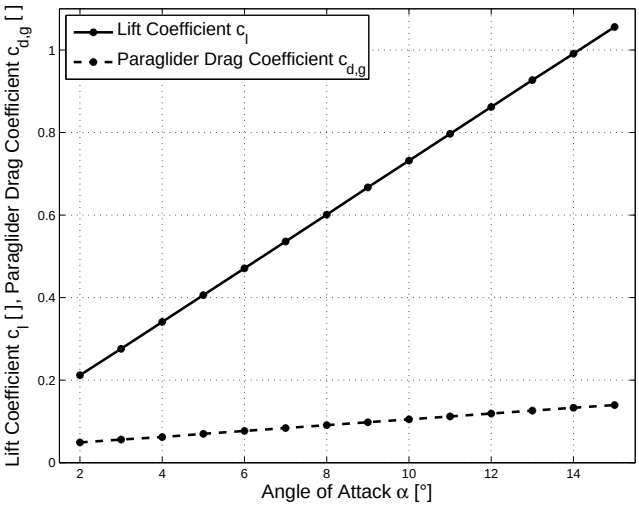


Figure 8. Aerodynamic coefficients as functions of α.

4 Simulation Results

The model is implemented in Matlab/Simulink. Two different flight situations are simulated. The first one is a gliding flight, where the pilot begins to fly at a certain height. The second one extends the first model by a winch launch to enable the pilot to gain altitude.

4.1 Gliding Flight

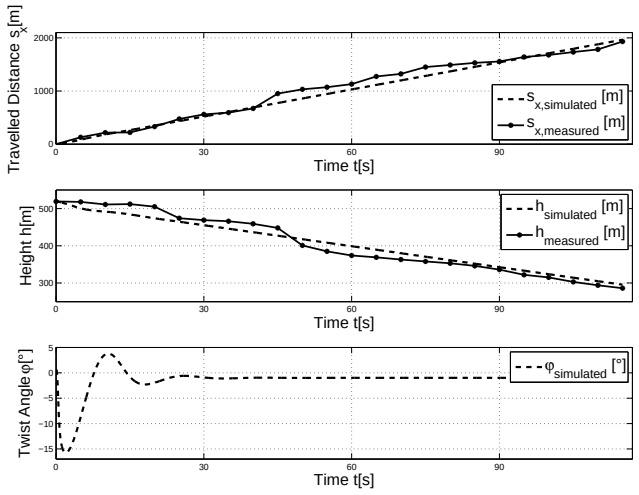


Figure 9. Gliding flight.

The three plots seen on Figure 9 show the translational motion of the system over the time. The angle of glide β is constant during the whole flight. With a start height h of 500 metres, a flight distance in x-direction s_x of 1930 m is reached on the expense of a 230 m loss in altitude. The third plot illustrates the dynamics of the twist angle φ with a settling time of about 30 s. The deviation between the simulated and measured results may be due to different wind situations. The wind affects the system behaviour significantly (see Figure 10). At the beginning,

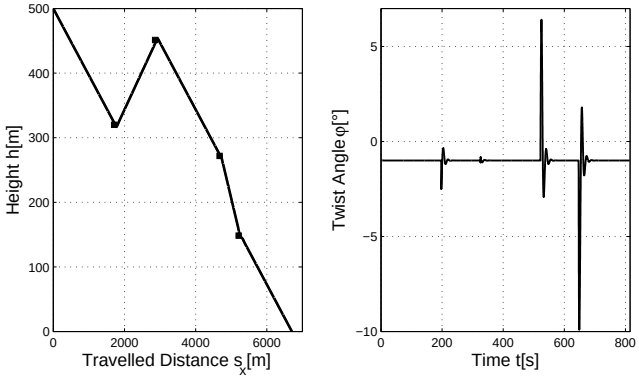


Figure 10. Gliding flight with wind impact.

the pilot glides down and is slowly reducing altitude. Subsequently, the pilot enters an updraft region (first black square) and, as a result, the pilot gains altitude and rises up to 450 m (second black square), where the helpful wind stops. The pilot glides down to the third black square, where headwind prevails. Consequently, the ground speed v_{GS} reduces and the angle of glide β increases. Therefore, height is lost more quickly than before and, as a result, the travelled flight distance s_x gets shorter. At a height of 150 m (fourth black square), the headwind levels off and the paraglider flies again with trim speed. The reaction of the twist angle φ in relation to the changing conditions is displayed on the right side of Figure 10. If the ground speed v_{GS} changes due to the impact of the wind, the angle of twist φ is disturbed from its steady-state value and after short time it is settled again. The disturbance caused by horizontal wind is stronger than the vertical wind, because headwind or tailwind directly creates a torque around the gravity point G and deflects the system from the rest position.

4.2 Towing Process

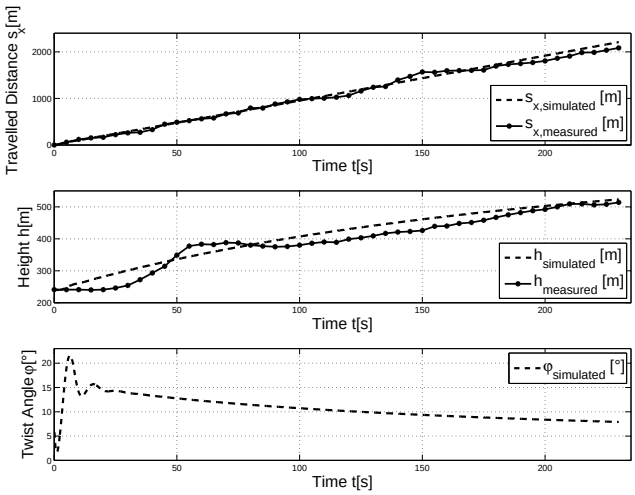


Figure 11. Towing process.

The first diagram of Figure 11 shows the travelled distance of the paraglider in x-direction s_x over the time. It is easy to recognize that the movement is linear due to the constant slope of the graph. The second plot shows the height h, whereas the transient response of the twist angle φ is drawn in the last diagram of Figure 11. A comparison of measured and simulated results reveals that the simulation model provides a fair reflection of the reality.

5 Conclusions

A simulation model for a paraglider is derived and implemented in Matlab/Simulink. The parameters are adjusted in order to simulate the behaviour of a certain paraglider. With the help of the model two different flight situations, i.e. gliding flight and towing process, are simulated. Simulation results are compared with data of a real flight. The model provides a good approximation of the dynamical behaviour of the real system. It allows to compare different parameter settings, determine their influence on the system as well as to optimise the towing process. However, as already mentioned, paragliding is under constant change and this model has to be adjusted accordingly.

References

J. D. Anderson. *Introduction to Flight*. McGraw-Hill, New York, 5th edition, 2005.

I. Currer. *Touching Cloudbase - The Complete Guide to Paragliding*. Air Supplies, York, 5th edition, 2011.

J. Dankert and H. Dankert. *Technische Mechanik: Statik, Festigkeitslehre, Kinematik/Kinetik*. Vieweg+Teubner, Wiesbaden, 5th edition, 2009.

H. Fahr. *Gleitsegelschlepp*. Flieg Zeug, Magdeburg, 2nd edition, 1992.

P. Janssen, K. Slezak, and K. Tänzler. *Gleitschirmfliegen: Theorie und Praxis*. Nymphenburger, München, 18th edition, 2013.

F. Karbstein. *Richtig Paragliding*. BLV Verlagsgesellschaft, München, 1996.

H. Oertel. *Introduction to Fluid Mechanics: Fundamentals and Applications*. Universitätsverlag, Karlsruhe, 2005.

O. Voigt. *Aerodynamik und Flugmechanik des Gleitschirms*. Books on Demand (Schweiz), Norderstedt, 2nd edition, 2003.

N. Whittall. *Paragliding: The Complete Guide*. Lyons Press Series. Lyons Press, Lanham, 1995.

N. Whittall. *Hang Gliding & Paragliding*. Clash series. Ticktock Media, Kent, 2010.

Modelling of a New Compton Imaging Modality for an In-Depth Characterisation of Flat Heritage Objects

Patricio Guerrero[1,2,3] Mai K. Nguyen[2] Laurent Dumas[3] Serge X. Cohen[1]

[1]IPANEMA USR 3461, CNRS/MCC/UVSQ/MNHN, Gif-sur-Yvette, France,
{patricio.guerrero-prado,serge.cohen}@university.org
[2]Equipes de Traitement de l'Information et Systèmes UMR 8051, ENSEA/UCP/CNRS
mai.nguyen-verger@u-cergy.fr
[3]Laboratoire de Mathématiques de Versailles UMR 8100, UVSQ/CNRS, Versailles, France,
laurent.dumas@uvsq.fr

Abstract

Objects having a flattened geometry, such as those encountered in heritage, have always been difficult to be analysed with conventional X-ray tomography methods due to their anisotropic morphology. To overcome the limitations of classical tomography for such samples, we envisage a new imaging modality based on Compton scattering. While Compton effect is usually considered as noise in tomography, in Compton scattering tomography the conditions are set such that this becomes the imaging agent of the image formation process. Our interests are, firstly, to avoid the relative rotation between the object, the source and the detector, and secondly, to be able to obtain in-depth data even when the sample is supported by some deep or dense material by exploiting only back-scattered photons. To replace the information provided by multiple projections angles in classical tomography, we make use of the relation between the energy loss of the scattered photons and its scattering angle, the Compton equation. Modelling of this new modality, image formation and object reconstruction through a filtered back-projection algorithm of a Radon transform on a half-space is presented. The feasibility of this concept is supported by numerical simulations.

Keywords: Radon transform, Compton scattering tomography, image reconstruction

1 Introduction

1.1 Context

Analysis of flat objects has remained particularly challenging with X-ray imaging methods. X-ray imaging has the strong advantage of providing non-invasive/non-destructive probing of the material and enables 2D two-dimensional (2D) imaging of the sample which is often mandatory when analysing heritage samples. However, a large set of those samples has a complex flattened geometry, *i.e.*, a large ratio between the front area and its thickness and the high differential light path in distinct directions prohibits the relative rotations of the sample to perform a three-dimensional reconstruction as would be done in regular tomography, using either absorption or phase contrast modality. Such samples could be encountered in conservation/restoration studies of easel painting requiring the characterisation of the stratigraphical assemblage of pigments often over a very dense background layer (Figure 1), *e.g.* made of white lead. Another type of samples falling in this set is encountered in palaeontology with the Lagerstätten fossils (Gueriau et al., 2014) which are mechanically flattened during the fossilisation process and stand on one side of a thick sedimental slab which can not be thinned for the study (Figure 2). In both cases the volume of interest forms a layer on top of a material matrix which is opaque to X-ray either due to its density or its thickness.

In both cases scientists left with one of two choices : either to perform a stratigraphical section of the sample which is an invasive method, or perform a 2D analysis, for example with synchrotron X-ray fluorescence spectral raster-scanning (Gueriau et al., 2014) while an in-depth reconstruction is still desired. Furthermore, because of the opaque supporting material, transmission and forward scattering data, with a scattering angle inferior of $\frac{\pi}{2}$, are impossible to collect. This leads us to propose a modality tapping on back-scattered data, that is, data collected with a scattering angle comprised between $\frac{\pi}{2}$ and π, as shown in Figure 3.

Using the approximation that all electrons of the material are both free and at rest, the Compton equation establishes in (1) a diffeomorphism between the scattering angle and the scattered energy, provided that the incident beam is monochromatic, *i.e.* a single value E_0 of an incident energy is illuminating the object. In such circumstances the Compton scattering image formation process corresponds to a Radon transform over either one branch of a V-line or equivalently over lines in a half-space (Morvidone et al., 2010; Truong and Nguyen, 2011, 2015) when the scene is purely 2-dimensional, or over the surface of a cone (Nguyen et al., 2005; Cebeiro et al., 2015) when a 3-dimensional scene is considered.

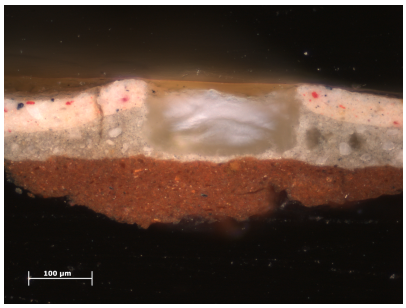


Figure 1. Paint cross-section showing a stratigraphical assemblage of *The Anatomy Lesson of Dr. Nicolaes Tulp*, 1632 by Rembrandt, Mauritshuis, The Hague. © Sample taken and prepared as cross-section by P. Noble during the conservation treatment of the painting in 1997, and re-photographed by A. van Loon, Mauritshuis, in 2010 for the Rembrandt Database.

Figure 2. A flat fossil actinopterygian over a high thick X-ray absorbing support from the Kem Kem Beds in Morocco dated back to the Lower Cretaceous (95 million years ago). © P. Gueriau (MHNM/MNHN).

1.2 Compton scattering tomography (CST)

Three basic photon-matter interaction phenomena are considered in classical X-ray imaging and hence tomography : Photoelectric absorption, Rayleigh scattering which is both elastic and coherent, and Compton scattering which conversely is both inelastic and incoherent. In this last one, an incident photon of energy E_0 is absorbed by a target electron, who re-emits a secondary photon scattered by an angle ω relative to the direction of the original photon. The scattered photon has then an energy E_ω which is related to the scattering angle ω by (1), the Compton equation, when the target electron is both free and at rest.

In classical X-ray imaging and tomography, Compton scattered signal is considered as noise added to photoelectric absorption and coherent scattered data. This is because X-ray transmission signal is dominated by the photoelectric absorption whilst coherent scattering may produce significant amplitude variations at low scattering angles thanks to constructive and destructive interference effects due to the coherent nature of this scattering. However, depending on the material, if the incident radiation has an energy superior of $40-50$ keV, Compton scattering is becoming the dominant phenomenon in the process, even more when detection is performed far from the transmission geometry ($\omega = 0$).

Classical tomographic imaging modalities, developed and used in most all applications in the last half century include : Transmission Computed Tomography (CT), Single Photon Emission Tomography (SPECT) and Positron Emission Tomography (PET). All of them regard primary radiation and perform a 3D mapping leaning on relative rotations between the object and the imaging setup. In such framework the Compton scattering is adding a non-uniform background to the observation, a systematic bias which leads to artefacts in the reconstruction if it is not accounted for. As the relative importance of the Compton scattering over the other two processes mentioned above is increasing with an increase on incident energy, its effect is even more important when using higher energy γ-ray imaging and tomography.

The idea of exploiting scattered radiation by Compton effect in imaging techniques has been introduced and studied simultaneously and it has given birth to CST (Norton, 1994; Nguyen et al., 2005; Morvidone et al., 2010; Cebeiro et al., 2015; Truong and Nguyen, 2011, 2015), which focuses to reconstruct the electron density map of the object.

To replace the information provided by multiple projections angles in classical transmission or emission tomography, CST exploits the energy loss of the scattered photons. This energy loss encodes the scattering angle information thanks to the Compton equation given by

$$E_\omega = \frac{E_0}{1+\frac{E_0}{m_e c^2}(1-\cos\omega)}, \tag{1}$$

where $m_e c^2 = 511$ keV is the rest mass energy of the electron.

The first CST scanner was proposed in 1994 (Norton, 1994) through a Radon transform over arc of circles starting at a γ-ray source and ending at the detecting point. Radon transforms over V-lines and its analytic inversion formula as well as a filtered back-projection reconstruction was proposed in (Morvidone et al., 2010), three different types of these V-line transforms with fixed axis direction are studied in (Truong and Nguyen, 2011) and new properties and applications including a Radon transform on a half-space were presented recently in (Truong and Nguyen, 2015). Radon transforms over conical surfaces having fixed axis directions and variable opening angle are studied in (Nguyen et al., 2005) where an analytic inversion formula is proposed, with applications to emission imaging based on Compton scattered radiation. A back-projection inversion for this conical Radon transform was developed recently in (Cebeiro et al., 2015).

This paper is organised as follows. Section 2 describes this new CST modality, and presents image formation process as well as object reconstructions via a back-projection algorithm related to a half-space Radon transform in the framework of a two dimensional setup with back-scattering. In Section 3 we present numerical scheme and some simulation results for flat objects both for energy resolved image formation and sample reconstruction. Finally, Section 4 closes this paper with a conclusion and perspectives.

The three dimensional case, corresponding to the conical Radon transform, will not be presented due to fact that the azimuthal scattering angle is uniformly distributed in a Compton event, that is to say, the angular distribution of scattered photons has axial symmetry around the direction of incidence.

2 Modelling of the new CST modality

CST focuses to reconstruct the electron density map of the object. In this paper, this density will be represented by a nonnegative bump function (both smooth and of bounded support) $f : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}^+$.

Fundamentally, an incident photon of energy E_0, coming from a point source **S**, is inelastically scattered by an electron located inside the object at **M** subtending an angle ω ($\frac{\pi}{2} < \omega < \pi$) with the direction of incidence. The scattered photon of energy E_ω, approximated by (1), arrives finally at a detecting point **D**. In this paper, **D** is located over the Oy-axis of the Cartesian reference plane and noted $\mathbf{D}_\zeta = (0, \zeta)$. Obviously we will have data for values of ζ verifying $|\zeta| > \zeta_0$ for some $\zeta_0 > 0$. The parallel beam is centred at the Ox-axis. See Figs. 3 and 4 for more details.

As already mentioned, in this paper we will considerer a two dimensional setup, *i.e.*, a slice of a three dimensional space through the z-axis.

2.1 Proposed setup by back-scattering

A synchrotron radiation setup with a monochromatic parallel X-ray beam (about 50 keV) and a space-energy resolved detector is proposed. As represented in Figure 3, the detector will be placed between the source and the object to capture back-scattered photons. It will have then a hole in the middle of length $2\zeta_0$ to allow the beam to go through. Vertical translations of the sample will allow the imaging of the full sample despite the limited beam size.

This configuration is considered to overcome limitations of getting data through classical methods mentioned before.

In Figure 4, scattered radiation arriving to a detecting point $\mathbf{D}_\zeta$ laying on the line that crosses both $\mathbf{M}_1$ and $\mathbf{D}_\zeta$ will have an energy loss related to ω_1, for example radiation scattered from $\mathbf{M}_3$. The same will occur related to $\mathbf{M}_2$ and ω_2. One will be able to identify, for example, points $\mathbf{M}_1$ and $\mathbf{M}_2$ radiated with the same X-ray beam at the same time through energy loss data.

2.2 Direct problem : Image formation

2.2.1 The half-space Radon transform (HRT)

The half-space Radon transform and its analytic inverse formula were introduced in (Truong and Nguyen, 2015). Briefly, it is a Radon transform, introduced in (Radon, 1917), defined on half-lines, starting on a detecting point towards the direction of where the sample is placed. The interest is that it performs an integration of the object, in this case a function representing an electron density distribution, over all locations where photons are scattered by the same angle and arriving at the same detecting point.

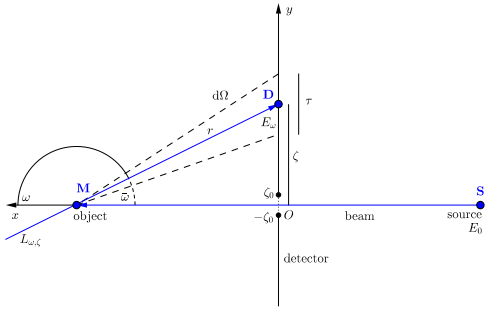


Figure 3. Simulated setup. Photons coming from a source at **S** are back-scattered at **M** by an angle ω into the solid angle dΩ. They are finally detected at **D**, who lays on the Oy-axis.

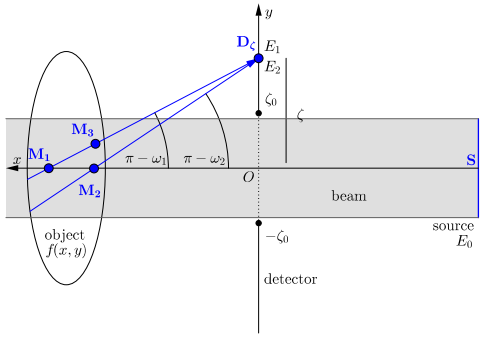


Figure 4. Three positions inside the sample producing scattered radiation captured at a detecting site $\mathbf{D}_\zeta$ at different energies $\{E_1, E_2\}$. $\mathbf{M}_1$ and $\mathbf{M}_3$ having the same scattering angle, generate both energy E_1, $\mathbf{M}_2$ does the same for energy E_2.

In a forward scattering framework, integration must be performed over a pair of half lines in the shape of a V, with fixed axis direction, these so-called V-line Radon transforms were introduced in (Morvidone et al., 2010) and developed in (Truong and Nguyen, 2011, 2015).

For simplicity, call $\bar{\omega} = \pi - \omega$ the supplementary angle of ω. To cover all projections from the sample to the full detector length, we have to consider $\frac{\pi}{2} < \omega < \frac{3\pi}{2}$ even if intervals $[\frac{\pi}{2}, \pi]$ and $[\pi, \frac{3\pi}{2}]$ are physically equivalents. Therefore, we will have $-\frac{\pi}{2} < \bar{\omega} < \frac{\pi}{2}$.

The half-space Radon transform, represented by R^*, is then defined in this paper, for $-\infty < \zeta < \infty$ and $-\frac{\pi}{2} < \bar{\omega} < \frac{\pi}{2}$, as the integral

$$R^* f(\bar{\omega}, \zeta) = \int_0^\infty \frac{1}{r} f(r\cos\bar{\omega}, \zeta - r\sin\bar{\omega})\,\mathrm{d}r. \qquad (2)$$

The factor $\frac{1}{r}$ in the last equation comes from the photometric law or the difference in solid angle from the scattering site to the detecting site. dr is the measure on the half-line.

2.2.2 Image formation

Let $\mathrm{d}I(E_\omega, \mathbf{D}_\zeta)$ represent the recorded scattered photon flux time density. It is the number of photons of energy laying in $E_\omega \mathrm{d}\omega$ recorded per unit time at $\mathbf{D}_\zeta$. It will incorporate the following parameters :

- I_0: the incident photon flux density before the scattering event.
- $\sigma^{2\mathrm{D}}(\omega)$: the 2D Klein-Nishina differential cross-section (Klein and Nishina, 1928) at an angle ω.
- $f(\mathbf{M})$: the electron density at $\mathbf{M}$.
- $\mathrm{d}\Omega(\mathbf{M}, \mathbf{D}_\zeta)$: the solid angle from $\mathbf{M}$ to $\mathbf{D}_\zeta$.
- $\mathrm{d}\mathbf{M}$: the elementary length around $\mathbf{M}$ over the half-line.
- $\mathrm{d}\omega$: the elementary variation of ω.

The solid angle $\mathrm{d}\Omega(\mathbf{M}, \mathbf{D}_\zeta)$ can be seen from Figure 3 to be

$$\mathrm{d}\Omega(\mathbf{M}, \mathbf{D}_\zeta) = 2 \arctan\left(\frac{\tau}{2r}\cos\bar{\omega}\right), \tag{3}$$

where τ is the length of the detecting element located at $\mathbf{D}$ and r the Euclidean distance from $\mathbf{M}$ to $\mathbf{D}$.

If τ is small enough, then $\mathrm{d}\Omega(\mathbf{M}, \mathbf{D}_\zeta)$ can be approximated by $\frac{1}{r}\tau\cos\bar{\omega}$.

The 2D Klein-Nishina differential cross-section is a function of the scattering angle ω giving the probability of a photon to be scattered by ω limited to a bi-dimensional scatter plane. It is given by

$$\sigma^{2\mathrm{D}}(\omega) = \frac{1}{2}\pi r_e^2 \left(\frac{E_\omega}{E_0}\right)^2 \left(\frac{E_\omega}{E_0} + \frac{E_0}{E_\omega} - \sin^2\omega\right), \tag{4}$$

where r_e is the classical electron radius.

Consequently, the scattered photon flux density at $\mathbf{D}_\zeta$, given a scattering site $\mathbf{M}$, is given by

$$\mathrm{d}I(E_\omega, \mathbf{D}_\zeta|\mathbf{M}) = I_0\, \sigma^{2\mathrm{D}}(\omega)\, \mathrm{d}\Omega(\mathbf{M}, \mathbf{D}_\zeta)\, f(\mathbf{M})\, \mathrm{d}\omega\, \mathrm{d}\mathbf{M}. \tag{5}$$

The recorded scattered flux time density $\mathrm{d}I(E_\omega, \mathbf{D}_\zeta)$ recorded at $\mathbf{D}_\zeta$ is the integral over all scattering sites laying on the half-line that starts at $\mathbf{D}_\zeta$ towards the object with slope $\tan\bar{\omega}$ noted $L_{\omega,\zeta}$. It is hence given by the integral

$$\mathrm{d}I(E_\omega, \mathbf{D}_\zeta) = \int_{\mathbf{M}\in L_{\omega,\zeta}} \mathrm{d}I(E_\omega, \mathbf{D}_\zeta|\mathbf{M}). \tag{6}$$

From the last integral, we can extract the half-space Radon transform, and we are able to express the scattered photon flux time density for photons laying in a recorded energy $E_\omega \mathrm{d}\omega$ as

$$\mathrm{d}I(E_\omega, \mathbf{D}_\zeta) = \tau I_0 \cos\bar{\omega}\, \sigma^{2\mathrm{D}}(\omega) R^* f(\bar{\omega}, \zeta)\, \mathrm{d}\omega. \tag{7}$$

2.3 Inverse problem : Object Reconstruction

2.3.1 Inversion of the HRT

In order to obtain an inverse formula for (2), we make use of Fourier techniques developed in (Morvidone et al., 2010) related to the V-line Radon transform.

Let $f(x,y)$ be expressed using its y-Fourier transform, noted $\hat{f}(x,q)$, as

$$f(x,y) = \int_{-\infty}^{\infty} \hat{f}(x,q) \mathrm{e}^{2\pi i y q}\, \mathrm{d}q. \tag{8}$$

Then (2) takes the form

$$R^* f(\bar{\omega}, \zeta) = \int_0^{\infty} \frac{1}{r} \int_{-\infty}^{\infty} \hat{f}(r\cos\bar{\omega}, q)\, \mathrm{e}^{2\pi i q(\zeta - r\sin\bar{\omega})}\, \mathrm{d}q\, \mathrm{d}r, \tag{9}$$

and applying Fubini's theorem,

$$R^* f(\bar{\omega}, \zeta) = \int_{-\infty}^{\infty} \mathrm{e}^{2\pi i q\zeta} \int_0^{\infty} \frac{1}{r} \hat{f}(r\cos\bar{\omega}, q)\, \mathrm{e}^{-2\pi i q r\sin\bar{\omega}}\, \mathrm{d}r\, \mathrm{d}q, \tag{10}$$

from where we can extract the ζ-Fourier transform of $R^* f(\bar{\omega}, \zeta)$ by writing

$$\widehat{R^* f}(\bar{\omega}, q) = \int_0^{\infty} \frac{1}{r} \hat{f}(r\cos\bar{\omega}, q) \mathrm{e}^{-2\pi i r q \sin\bar{\omega}}\, \mathrm{d}r. \tag{11}$$

With a change of variables $x = r\cos\bar{\omega}$, $t = \tan\bar{\omega}$ and defining the function $g : (t,q) \mapsto \widehat{R^* f}(\bar{\omega}, q)$, (11) becomes

$$g(t,q) = \int_0^{\infty} \frac{1}{x} \hat{f}(x,q) \mathrm{e}^{-2\pi i q t x}\, \mathrm{d}x, \tag{12}$$

which is the Fourier transform of the function $x \mapsto \frac{1}{x}\hat{f}(x,q)$ evaluated at qt. One can now apply the inverse formula and the scaling property of Fourier transforms to get $\hat{f}(x,q)$ in the form

$$\hat{f}(x,q) = |q| x \int_{-\infty}^{\infty} g(t,q) \mathrm{e}^{2\pi i q t x}\, \mathrm{d}t. \tag{13}$$

Finally, $f(x,y)$ can be reconstructed for all points (x,y) in the object through the q-inverse Fourier transform of the last expression.

2.3.2 Filtered back-projection

As the classical Radon Transform on straight lines, the filtered back-projection inversion on on a half-space combines the back-projection operation and a filtering operation. It says that the back-projection operator $b_\omega(x,y)$ at

(x,y) for an scattering angle ω assigns to (x,y) the value of the projection recorded at $\zeta = y + x\tan\bar{\omega}$ where (x,y) was projected, recalling that $\bar{\omega} = \pi - \omega$. That is to say, the back-projection operator is defined by

$$b_\omega(x,y) = R^* f(\bar{\omega}, y + x\tan\bar{\omega}). \tag{14}$$

A first rough reconstruction of the object can be done by applying the back-projection operator without filtering. The object can then be reconstructed by

$$f(x,y) = \int_0^{\pi/2} b_\omega(x,y)\,\mathrm{d}\bar{\omega}. \tag{15}$$

The filtering operation is applied to the ζ-Fourier transform of $R^* f(\bar{\omega},\zeta)$. To obtain the correct filter, we can write the inversion formula of (2) from (8), (13) and the Fubini's theorem as

$$f(x,y) = x\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |q| g(t,q) \mathrm{e}^{2\pi i q(y+xt)}\,\mathrm{d}q\,\mathrm{d}t, \tag{16}$$

where we identify the inverse Fourier transform of the function $q \mapsto |q|g(t,q)$ evaluated at $y+xt$ by writing

$$f(x,y) = x\int_{-\infty}^{\infty} \mathscr{F}^{-1}(|q|g)(t, y+xt)\,\mathrm{d}t, \tag{17}$$

where $\mathscr{F}^{-1}$ is the inverse Fourier transform.

Finally, in terms of the angle $\bar{\omega}$, last inversion formula reads

$$f(x,y) = x\int_{-\pi/2}^{\pi/2} \mathscr{F}^{-1}(|q|\widehat{R^*f})(\bar{\omega}, y + x\tan\bar{\omega})\,\frac{\mathrm{d}\bar{\omega}}{\cos^2\bar{\omega}}. \tag{18}$$

The last inverse Fourier transform is called the filtered projection of R^*f using the so-called ramp filter $|q|$, it is represented by

$$\widetilde{R^*f}(\bar{\omega},\zeta) = \mathscr{F}^{-1}\left[|q|\widehat{R^*f}\right](\bar{\omega},\zeta), \tag{19}$$

and then one is able to write the back-projection operator related to this filtered projection as

$$\tilde{b}_\omega(x,y) = \widetilde{R^*f}(\bar{\omega}, y + x\tan\bar{\omega}), \tag{20}$$

to finally be able to write a filtered back-projection reconstruction of $f(x,y)$ for all (x,y) in the object under a compact form via the integral

$$f(x,y) = x\int_{-\pi/2}^{\pi/2} \tilde{b}_\omega(x,y)\,\frac{\mathrm{d}\bar{\omega}}{\cos^2\bar{\omega}}. \tag{21}$$

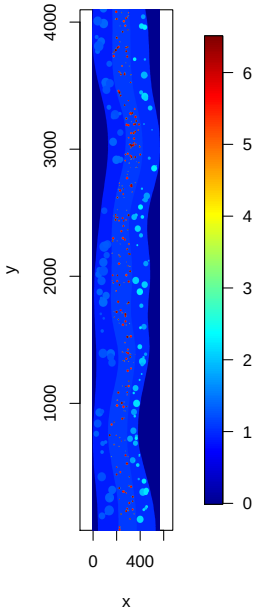


Figure 5. Test sample: A flat stratigraphic phantom.

3 Simulation of the new CST modality

A numerical phantom corresponding to a transversal section of a flattened stratigraphic sample is generated and reconstructed in Figs. 5 and 6 respectively via inversion formula (21).

The phantom presents three layers of about 150×4096 square micrometers with electron densities of 0.9, 1.1 and 1.0 respectively. Grains of random diameter and position were inserted inside the layers, having densities of $1.3 \pm .1$, 6.0 ± 0.5 and 2.0 ± 0.3 respectively.

In these first attempts, attenuation and multiple scattering are neglected.

3.1 Energy resolution of the detector

Simulation of this new CST modality requires to consider a realistic energy-resolved detector. Let ΔE be the energy resolution of the detector. Hence, for a fixed detector, we will have a limited number of energy channels recording photons laying on a specific energy range.

Let C_i be the energy channel detecting all photons laying on the energy range $[E_i, E_i + \Delta E]$. It can be equivalently defined in terms of the scattering angle through (1) for ω laying in $C_i = [\omega_i, \omega_{i+1}]$.

The measured intensity into a given energy channel for a given detecting site is then given by integrating the scattered intensity over that channel

$$I(C_i, \mathbf{D}_\zeta) = \int_{C_i} I(E, \mathbf{D}_\zeta)\,\mathrm{d}E, \tag{22}$$

where $\mathrm{d}I(E,\mathbf{D}_\zeta)$ is given in (7).

The trapezoidal rule was used to compute numerically last integral with an angular step corresponding to 5 eV by means of (1). Consequently, the same parameters are used for integral (21) with a simple change of variables from ω to E through (1). Two energy resolutions are considered in simulations, namely $\Delta E \in \{50, 100\}$ eV.

3.2 Spatial discretization

The unit length considered is 2 µm. Each pixel in image representations will represent then 2×2 µm^2.

A parallel X-ray beam of 8 µm of width crosses a hole in the detector of 12 µm of width, hence we have $\zeta_0 = 6$ µm. The medium considered has 512×4096 µm^2 and then we need 512 vertical translations due to the height of the sample to cover the full phantom. The considered detector located at 2 µm of the sample over the Oy-axis has a vertical length of 2048 µm including the hole. Thereby, we have $1024 - 24$ detecting sites. (24 sites not considered due to the hole.)

Figure 7 shows the bottom slice of the sample (8 µm of heigh corresponding to the beam width) and its associated image formation through a discretization of (2).

Numerical resolution of integral (2) is performed via the trapezoidal rule with a discretization spatial step $\mathrm{d}r = 2$ µm.

3.3 Window functions

Window functions $\mathrm{w}(q)$ are added to the ramp filter in numerical reconstructions to control high frequencies. Filtered projections related to R^*f written in (19) are then rewritten in the form

$$\widetilde{R^*f}(\bar{\omega}, \zeta) = \mathscr{F}^{-1}\left[|q|\mathrm{w}(q)\widehat{R^*f}\right](\bar{\omega}, \zeta). \qquad (23)$$

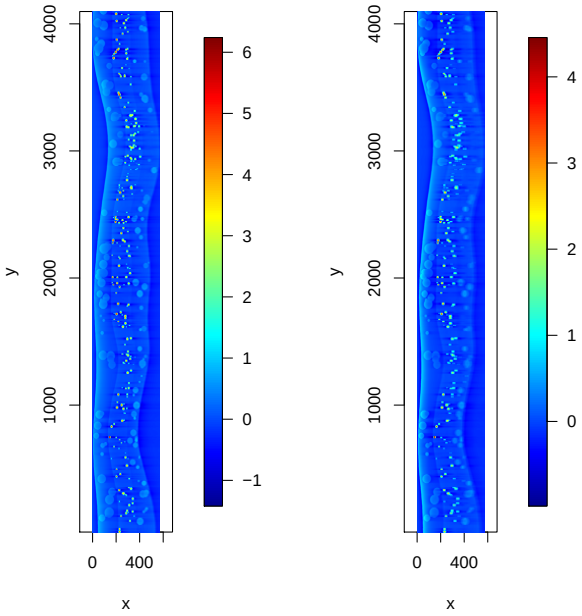
Cosine related window functions are widely used, they have the form

$$\mathrm{w}(q) = \cos^n(\pi q). \qquad (24)$$

Within this family of windows, the case $n = 2$ corresponds to the well characterised Hann Window function. During our simulation work we tested various values of n and found out that $n = 8$ was the one minimising the mean square error between the object and its reconstruction given the numerical noise produced by the simulation. Still different values of n gives satisfactory reconstructions as well. Horizontal artefacts in reconstructions are due to vertical translations of the object. Grains inside the matrix layers are reconstructed at the correct depth, size and relative densities, without the need of a relative rotation between phantom and detector.

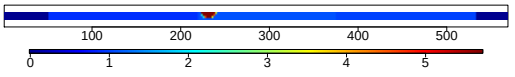
4 Conclusions and perspectives

A new X-ray imaging modality based on Compton scattering is proposed, the particular aim is a 3D reconstruction of flat objects without relaying on a relative rotation between the studied sample and the imaging setup. Modelling of both image formation by means of the half-space

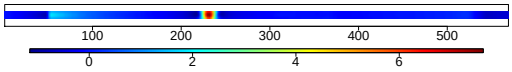


(a) Reconstruction with $\Delta E = 50$ eV. **(b)** Reconstruction with $\Delta E = 100$ eV.

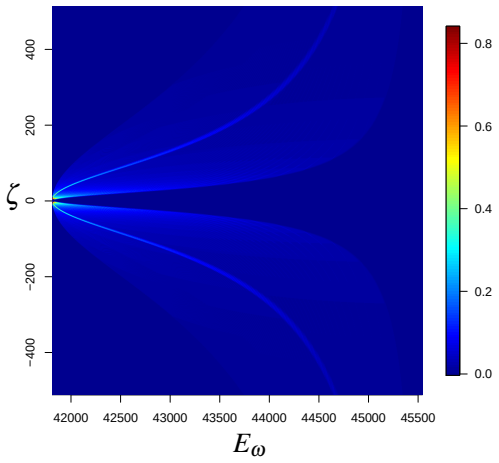
Figure 6. Numerical results of reconstructions from backscattered data.



(a) Bottom 8 µm of the phantom.



(b) Reconstruction of Figure 7a with $\Delta E = 50$.



(c) HRT of object shown in Figure 7a. We can recognise the effect of the middle red grain at the two highlighted curves.

Figure 7. Image formation process through this new modality.

Radon transform and object reconstruction by a filtered back-projection inversion are presented, considering realistic energy resolved detectors, and supported with numerical simulations proving feasibility with experimental data. Simulations of the three-dimensional setup, regarding a conical Radon transform, are in progress for both forward and backward scattering. Yet the present results are already very encouraging considering the problem of non-destructive and non-invasive 3D imaging of samples supported by a deep or dense material.

Acknowledgment

The authors would like to thank Pierre Gueriau and Mathieu Thoury for their support, discussions and help to acquire images of heritage objects. Annelies van Loon for providing the image of the Rembrandt painting, and the field workers who collected the fossil actinopterygian.

Patricio Guerrero would also like to thank the *Fondation des Sciences du Patrimoine* for providing a PhD grant to support his research work.

References

J. Cebeiro, M. Morvidone, and M. K. Nguyen. Back-projection inversion of a conical Radon transform. *Inverse Problems in Sciences and Engineering*, 24(2):328–352, April 2015.

P. Gueriau, C. Mocuta, D. B. Dutheil, S. X. Cohen, D. Thiaudière, S. Charbonnier, G. Clément, and L. Bertrand. Trace elemental imaging of rare earth elements discriminates tissues at microscale in flat fossils. *Plos One*, 9(1):e86946, 2014.

O. Klein and Y. Nishina. The scattering of light by free electrons according to dirac's new relativistic dynamics. *Nature*, 122 (3072):398–399, 1928.

M. Morvidone, M. K. Nguyen, T. T. Truong, and H. Zaidi. On the V-line radon transform and its imaging applications. *International Journal of Biomedical Imaging*, 2010:11, 2010.

M. K. Nguyen, T. T. Truong, and P. Grangeat. Radon transforms on a class of cones with fixed axis direction. *Journal of Physics A: Mathematical and General*, 38(37):8003, 2005.

S. J. Norton. Compton scattering tomography. *Journal of Applied Physics*, 76(4):2007–2015, 1994.

J. Radon. Berichte über die verhandlungen der königlich sächsischen gesellschaft der wissenschaften zu leipzig. *Mathem. Phys. Klasse*, 69:262–266, 1917.

T. T. Truong and M. K. Nguyen. On new V-line radon transforms in $\mathbb{R}^2$ and their inversion. *Journal of Physics A: Mathematical and Theoretical*, 44(7):075206, 2011.

T. T. Truong and M. K. Nguyen. New properties of the V-line radon transform and their imaging applications. *Journal of Physics A: Mathematical and Theoretical*, 48(40):405204, 2015.

Analysis of Optimal Diesel-electric Powertrain Transients during a Tip-in Maneuver

Vaheed Nezhadali Lars Eriksson

Vehicular Systems Division, Electrical Engineering Department
Linköping University, SE-581 83 Linköping, Sweden {vaheed.nezhadali,lars.eriksson}@liu.se

Abstract

Optimal transients of a hybrid powertrain are calculated with the aim to give a smooth and time efficient acceleration. It is shown that there is a trade-off between time and driveline oscillations where high oscillations can be avoided by slightly longer acceleration time and proper control of the electrical and diesel power sources. During a low oscillation acceleration, there is still the possibility to reduce the amount of total consumed electrical and fuel energy. This is investigated by calculation of optimal controls during acceleration for a fixed time while penalizing the usage of energy in a low oscillation acceleration. The balance between electrical and diesel energy usage during the acceleration is also investigated. The results show that to avoid extreme transients by optimal control, a multidimensional formulation of the objective function including different properties should be considered.

Keywords: numerical optimal control, acceleration, vehicle jerk

Nomenclature

The nomenclature for the paper with subscripts and variables is given in Tables 1 and 2 respectively.

Table 1. Subscripts used for variables.

Index	Description	Index	Description
im	Intake manifold	em	Exhaust manifold
gen	Generator	wg	Wastegate
e	Engine	a	Air
ds	Drive shaft	v	Vehicle
$mf, conv$	Fuel conversion	m_f	Fuel mass
$loss$	Losses	tot	Total
$mech$	Mechanical	tc	Turbocharger
w	Wheel	gb	Gearbox
fd	Final drive	$resist$	Resistant forces
c	Compressor	ac	Air into cylinder
0	Initial	f	Final
r	Radius	gs	Genset

Table 2. Variables used in the paper.

Symbol	Description	Unit
x	State variable	-
u	Control input	-
θ	Angle	rad
t	Time	s
F	Force	N
R	Gas constant	N·m/kg·K
p	Pressure	Pa
T	Temperature	K
M	Torque	N·m
k	Stiffness coefficient	N·m/rad
b	Damping coefficient	N·m·s/rad
ω	Rotational speed	rad· s^{-2}
α	Rotational acceleration	rad/s
β	Road slope	rad
$m, \dot{m}$	Mass, Mass flow	kg, kg/s
P	Power	W
E	Energy	J
u_{mf}, u_{wg}, P_{gen}	Control signals	mg/cycle, -, W
J	Inertia	kg · m^2
ρ	Density	kg · m^{-3}
r	Radius	m
A	Vehicle frontal area	m^2
BSR	Blade speed ratio	-
λ	Air/fuel equivalence ratio	-
ϕ	Fuel/air equivalence ratio	-
i	Gear ratio	-
η	Efficiency	-
Π	Compression ratio	-
c	Constant coefficient	-
ψ	Electrical energy penalty coefficient	-
δ	Energy penalty coefficient	-
$(A/F)_s$	Stoichiometric Air to fuel ratio	-

1 Introduction

Hybridization of powertrains opens up new opportunities for faster and more efficient vehicle acceleration. With an electric power source assisting a diesel engine, there is an extra degree of freedom in powertrain control while the simultaneous control of the diesel and electric power sources becomes more complex.

Tip-in maneuver is referred to the situation where the driver suddenly asks for a fast vehicle acceleration by pressing accelerator pedal. This is a highly demanding and transient operation in a diesel-electric powertrain. The controls during this period can be optimized with respect to energy consumption or the operations time similar to (Sivertsson and Eriksson, 2012b), (Sivertsson and Eriksson, 2012a) and (Sivertsson and Eriksson, 2015b). Passenger comfort is also important when considering powertrain transients and can be accounted for by taking the driveline oscillations, referred to as Jerk, into account. The Jerk is also important considering its effects on the life length of driveline components, for more discussion see (Haj-Fraj and Pfeiffer, 2001) and (Haj-Fraj and Pfeiffer, 2002). In real world applications, not a single but all of these ob-

jectives are of importance and therefore it is desirable to obtain a compromise between these objectives by proper control of the powertrain.

The contribution of this paper is the development of a methodology for the calculation of efficient hybrid powertrain transients with the aim to obtain a compromise between time-energy-Jerk objectives during a tip-in maneuver. Numerical optimal control is used as an enabler for this where first the extreme transients obtained by improper objective function formulations are presented. Then the trade-off between time-Jerk and Jerk-energy are calculated. The problem is solved for different road slopes representing various loading scenarios. The analysis is extended by investigation of powertrain transients and the balance between usage of diesel and electric energy sources is analyzed.

2 Powertrain model

To enable optimal control problem (OCP) formulation, a model for the powertrain and driveline components is needed. The powertrain model representing a hybrid bus is comprised of a diesel engine and an electric motor/generator working in parallel. The dynamics are described by a mean value engine model (MVEM) and a model for generator efficiency in a validated diesel-electric powertrain (genset) model from (Sivertsson and Eriksson, 2014). The powertrain dynamics are described by four state variables as $\omega_e(t)$, $p_{im}(t)$, $p_{em}(t)$ and $\omega_{tc}(t)$. Two additional states describe the dynamics of the driveshaft twist angle $\theta_{ds}(t)$ and wheel speed $\omega_w(t)$. The model has three control inputs for injected fuel during each combustion cycle $u_{mf}(t)$, wastegate position $u_{wg}(t)$ and the the electric power of motor/generator $P_{gen}(t)$.

Dynamics of the four genset state variables are described by the following differential equations:

$$\frac{d\omega_e}{dt} = \frac{1}{J_{gs}}(M_{gs} - M_{gs,load}) \tag{1}$$

$$\frac{dp_{im}}{dt} = \frac{R_{im}T_{im}}{V_{im}}(\dot{m}_c - \dot{m}_{ac}) \tag{2}$$

$$\frac{dp_{em}}{dt} = \frac{R_{em}T_{em}}{V_{em}}(\dot{m}_{ac} + \dot{m}_f - \dot{m}_t - \dot{m}_{wg}) \tag{3}$$

$$\frac{d\omega_{tc}}{dt} = \frac{P_t\eta_{mech} - P_c}{\omega_{tc}J_{tc}} \tag{4}$$

The flexibilities in the driveline are lumped into one single flexibility in the driveshaft according to (Pettersson and Nielsen, 2000), and the torque transferred by the driveshaft is described using the stiffness and damping coefficients as follows:

$$M_{ds} = k_{ds}\,\theta_{ds} + b_{ds}\frac{d\theta_{ds}}{dt} \tag{5}$$

$$\frac{d\theta_{ds}}{dt} = \frac{\omega_e}{i_{gb}\,i_{fd}} - \omega_w \tag{6}$$

where (6) is used to describe the driveshaft deflection dynamics.

Considering rolling and aerodynamic resistances and gravitational force, as well as constant gearbox and final drive ratios, the wheel speed dynamics are calculated using Newton's second law of motion as follows:

$$\frac{d\omega_w}{dt} = \frac{M_{ds} - M_{resist}}{J_w + m_v r_w^2} \tag{7}$$

$$M_{resist} = 0.5\,\rho_{air}\,c_a\,A\,\omega_w^2 r_w^3 + m_v\,g\,r_w(c_r\cos(\alpha) + \sin(\alpha)) \tag{8}$$

The utilized electric and diesel energy are represented by the following integral states:

$$E_{gen} = \int_{t_0}^{t_f} P_{gen}\,dt \tag{9}$$

$$E_{m_f} = q_{hv}\int_{t_0}^{t_f} u_{mf}\,\omega_e\,n_{cyl}\frac{10^{-6}}{4\pi}\,dt \tag{10}$$

When formulating OCPs, the oscillations in the rotational speed of the transmission shaft is used to represent the driveline oscillation. These oscillations are referred to as Jerk that is defined as follows:

$$Jerk = \int_{t_0}^{t_f} \dot{\alpha}_{tr}^2\,dt \tag{11}$$

$$\alpha_{tr} = \frac{d\omega_e}{dt}\frac{1}{i_{gb}} \tag{12}$$

3 Problem formulation

In this section, first definition of the tip-in problem in terms of boundary conditions and constraints is described and then the objective function formulation for the OCPs are presented.

3.1 Tip-in problem constraints

3.1.1 Boundary conditions for the tip-in problem

The tip-in starts from a stationary operating condition at constant vehicle speed of 10 km/h and the final condition is that the speed should reach 15 km/h. The states and control inputs should remain within the allowed limits during the operation while the integral states and generator power are assumed to be zero at the beginning. All these can be summarized as:

$$\begin{cases} \omega_w(t_0) = \frac{\omega_e(t_0)}{i_{fd\times gb}} = \frac{10}{r_w}\frac{1}{3.6}, \quad \dot{x}(t_0) = 0, \\ \omega_w(t_f) = \frac{15}{r_w}\frac{1}{3.6}, \\ E_{gen}(t_0) = E_{m_f}(t_0) = P_{gen}(t_0) = 0, \\ u_{min} \le u \le u_{max},\ x_{min} \le x \le x_{max} \end{cases} \tag{13}$$

3.1.2 Path constraints during tip-in

The problem is solved for a hybrid bus where the maximum acceleration of 1 m/s^2 according to the limits in SORT (Standardised On-Road Test cycles (SORT), Last

accessed April 2016) are used as the highest allowed acceleration. The SORT standard is used in Europe to design on-road test cycles in order to measure fuel consumption of buses. There are also constraints regarding the turbocharger operation to avoid surge, and operational region for the turbine blade speed ratio. The maximum engine power is limited according to the maximum power curve at different engine speeds and finally, the air to fuel ratio should satisfy the smoke limit constraint λ_{min}. There is also a mechanical limit on how fast the wastegate can be actuated and the rate of change in generator power. These constraints are summarized as:

$$\begin{cases} \frac{d\omega_w}{dt} \times r_w < 1,\ \Pi_c \leq \Pi_{c,surge}, \\ BSR_{min} \leq BSR(x,u) \leq BSR_{max}, \\ P_e(x,u) \leq P_{e,max}(x),\ \frac{\dot{m}_{ac}}{\dot{m}_f}(A/F)_s \leq \frac{1}{\lambda_{min}}, \\ |\dot{u}_{wg}| \leq c_{wg},\ |\dot{P}_{gen}|/\omega_e \leq c_{gen} \end{cases} \tag{14}$$

3.2 Optimal control problem formulation

In analysis of powertrain dynamics during tip-in, one objective is to calculate the minimum time transients of the powertrain. For this, OCPs with objective function of the following form are solved:

$$\min_{(x,u)} \int_{t_0}^{t_f} dt \tag{15}$$

The trade-off between minimum time and minimum Jerk transients will be calculated by first calculating the shortest time via solving the minimum time problem, and then, minimizing the Jerk using a fixed t_f. The time is then increased step wise compared to the calculated minimum time duration. The OCP formulation in this case looks as follows:

$$\min_{(x,u)} \int_{t_0}^{t_{f,fix}} \text{Jerk}\ dt \tag{16}$$

The energy from fuel and electrical sources during the vehicle acceleration can be minimized solving for:

$$\min_{(x,u)} \int_{t_0}^{t_{f,fix}} (E_{m_f} + E_{gen})\, dt \tag{17}$$

In (Nezhadali and Eriksson, 2016) it is discussed that after minimizing the Jerk in a fixed time OCP, energy consumption minimization is the next dimension that can be analyzed for a low Jerk solution. This is done by calculating the Jerk optimal control transients including a penalty δ on energy consumption while using a fixed time. The penalty on the energy consumption is increased iteratively and the problem is solved several times to obtain the trade-off between Jerk and energy objectives. The objective function formulation for this case is:

$$\min_{(x,u)} \int_{t_0}^{t_{f,fix}} \text{Jerk} + \delta \times (E_{m_f} + E_{gen})\, dt \tag{18}$$

There is still another dimension to the optimization problem which is the balance between usage of diesel and electrical energy during the acceleration. To investigate how such balance affects the system transients, the energy consumption is reformulated including a penalty ψ on the electrical energy consumption. The objective function formulation then looks as:

$$\min_{(x,u)} \int_{t_0}^{t_{f,fix}} \text{Jerk} + \delta \times (E_{m_f} + \psi \times E_{gen})\, dt \tag{19}$$

The problem in (19) is solved with various combinations of δ and ψ penalties.

Finally, the different OCP formulations with the objective functions mentioned above become:

Objective function in (15) or (16) or (17) or (19)

subject to:

$$\dot{x} = f(x,u)$$

Constraints in (13) and (14)

3.3 Numerical solution of optimal control problems

To solve the formulated OCP in the previous section, a direct multiple shooting method using CasADi software package (Andersson, 2013) is used. The dynamics in each discretization interval are forward integrated using a 4 step Runge-Kutta integrator. After discretization of constraints, objective function and the dynamics, a nonlinear programming problem (NLP) is formulated and solved using IPOPT (Wächter and Biegler, 2006) to obtain the optimal controls and corresponding state transients.

To ensure that the solutions are not affected by the number of discretization intervals, the problem is solved with different values. It is seen that the transients remain unchanged for intervals close to and more than 300 so this is chosen as the number of discretization intervals.

4 Optimal control results

4.1 Extreme transients

To show the importance of finding a compromise between time, Jerk, and energy objectives, the state and control transients are first presented for extreme cases where only one of these are considered in the optimization. These transients are obtained by solving for (15), and (16) and (17) the latter two with final time locked $t_{f,fix} = 2$ s.

The min T transients, illustrated in Figure 1, are very oscillatory at the beginning for all controls and such control strategy would have severe negative impacts on the passenger comfort as well as the life length of genset and driveline components. For the bus to be able to smoothly continue its movement after reaching the final speed, the twist angle in the driveshaft should match the required torque and acceleration at wheels. In min T transients, due to the high deflection in the driveshaft, the vehicle speed even at the end of the acceleration is still increasing at a

high rate. A transition from this high acceleration to a low acceleration would be undesirable in terms of passenger comfort standards.

The minimum energy transients are less oscillatory but diesel engine power is not used for vehicle propulsion and all required power is provided only by the electric motor. The very low engine speed at the end of the minimum energy transients, increases the risk of engine stall when the diesel engine is going to take over the power production after the acceleration which makes the controls less applicable in real world applications.

For min Jerk transients, the undesirable non-smooth speed transition at the end time is similar to the min T case. The controls are less oscillatory compared to the min T case, but the bang-bang looking controls are what the manufacturers are less willing to implement because of the issues with component wear and durability accompanied with such control strategies.

Considering the mentioned drawbacks, these solutions are considered extreme and less applicable for control design in real world applications. In the following sections, the transients obtained by the suggested methodology for finding proper compromise between time-energy-Jerk objectives are presented and analyzed.

4.2 Compromise between time, Jerk and energy

Figure 2 shows the trade-off between time and Jerk objectives calculated by solving (16) as stated in Section 3.2 for three different road slopes. The Jerk in min T solution, calculated by solving (15), is extensively larger and therefore it is not included in the trade-offs. However, the Jerks in Figure 2 are normalized with respect to the largest Jerk belonging to the min T solution of the *0 degree slope* case which is referred to as $Jerk_{max}$. It is seen that the Jerk can be significantly decreased compared to the min T solution for all road slopes. It is desirable to have small Jerk during operation, specifically in a city bus application. Therefore, a duration of 2 [s] where the Jerk approach near zero values for all road slopes is chosen as the fixed time duration for which the energy-Jerk trade-off is calculated.

4.3 Jerk-Energy trade-off

To investigate the energy balance during the genset operation, different energy components and fuel conversion efficiency are defined as follows:

$$E_{tot} = E_{mf} + E_{gen} \tag{20}$$

$$E_{loss} = E_{mf} + E_e \tag{21}$$

$$E_e = \int_{t_0}^{t_f} M_e \omega_e \, dt \tag{22}$$

$$\eta_{mf,conv} = \frac{E_e}{E_{mf}} \tag{23}$$

where E_e represents the net energy from the diesel engine which is used for acceleration, and E_{loss} represents the losses such as engine friction and pumping work.

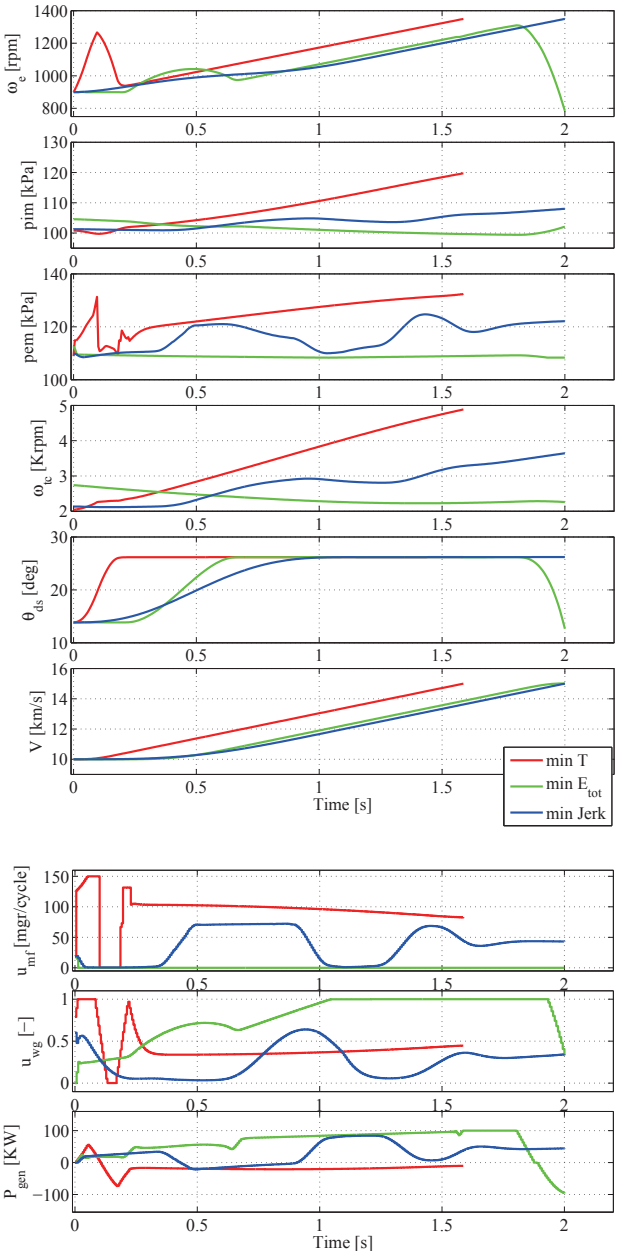


Figure 1. Optimal state and control transients for the extreme cases during acceleration ($E_{tot} = E_{mf} + E_{gen}$).

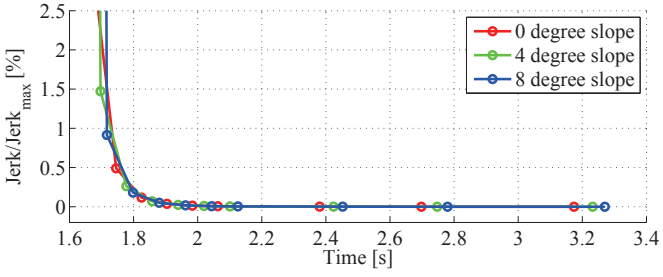


Figure 2. Trade-off between Jerk and time objectives.

To obtain the trade-off between Jerk and energy for different road slopes, (19) is solved while increasing δ and

ψ. The results for the three road slopes are presented in Figure 3. Reminding that the purpose of applying energy penalties is to avoid "extreme" low Jerk transients, a point with slightly increased Jerk on the trade-offs is chosen as a "candidate" case for which the energy balance is presented in Figure 4. Independent of what power is required for acceleration, the total required energy shows a decreasing trend when the penalty δ is increased. Also, when the penalty ψ is increased, meaning that the usage of electrical energy becomes more costly in the (19), more power is delivered by the diesel engine which has a low efficiency and therefore the total required energy for vehicle acceleration increases. As seen in Figure 4, for larger ψ values, less electrical energy is used and when total required energy for acceleration is low, the 0 degree case, diesel engine power is even used to produce electrical energy in addition to vehicle acceleration.

In case of the 0 degree slope, the required energy for the $\psi = 0$ remains unchanged for all δ values. This is because this operating condition requires smaller amount of energy compared to other cases while $\psi = 0$ in (19) implies that it does not have any cost to use electrical energy. Low efficiency of the diesel engine compared to the generator and the cheapness of electrical energy makes it optimal to perform the acceleration using only the electrical energy with no regard to the penalty δ on total energy consumption. This can be verified comparing the $\psi = 0$ energy balance for the three road slopes in Figure 4.

The first time diesel power is used for acceleration in the 0 degree slope is at $\psi \approx 2.05$. At this operating condition, a sudden decrease in E_{tot} takes place according to Figure 4. Moving from the first point on the $\psi = 2.05$ to the second point of the curve in Figure 3, the increase in δ makes the contribution from the energy term larger than the Jerk term in (19). As a result, a higher efficiency in energy consumption is favored. Since usage of the fuel energy accompanies high losses, achieving higher total efficiency is facilitated by altering the contribution of energy sources from very low electrical energy usage, similar to $\psi = 2.08$ in the 0 degree slope of Figure 4, into nearly equal contribution from the electrical and fuel energy sources, in $\psi = 2.05$.

Considering the fuel conversion efficiencies presented in Table 3, the efficiency is lower at low loads corresponding to the 0 degree slope and when electrical energy is cheaper to use (smaller ψ values). Other than this, an efficiency close to 40 % is maintained at different loading conditions.

Table 3. Fuel conversion efficiency.

road slope	ψ	0	2.05	2.08	2.5	4
0 [deg]	$\eta_{mf,conv}$[%]	-16.6	34.7	39.9	40.6	40.3
road slope	ψ	0	2.02	2.05	2.2	4
4 [deg]	$\eta_{mf,conv}$ [%]	28.74	38.02	40.16	40.86	40.13
road slope	ψ	0	2.1	2.2	2.5	4
8 [deg]	$\eta_{mf,conv}$ [%]	39.6	40.94	41.08	40.59	39.43

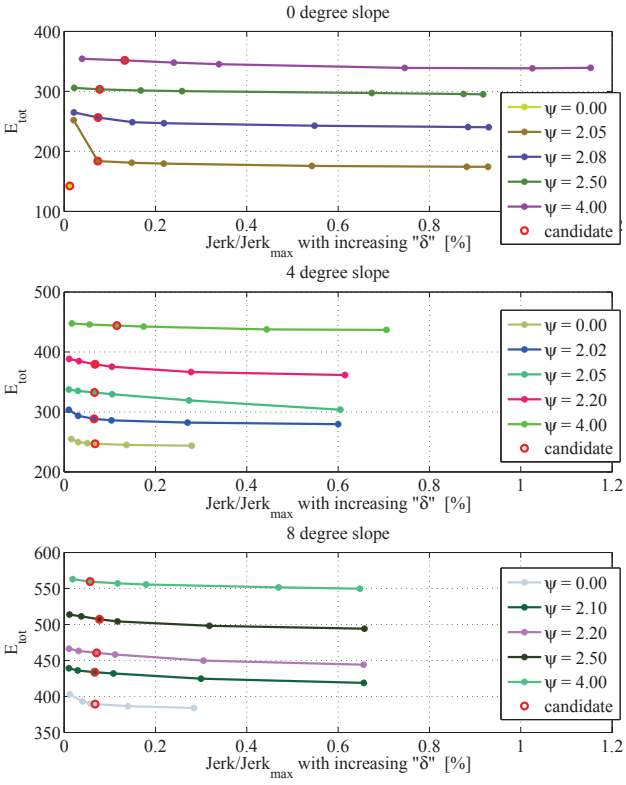


Figure 3. The trade-off between Jerk and energy with different energy penalties and road slopes.

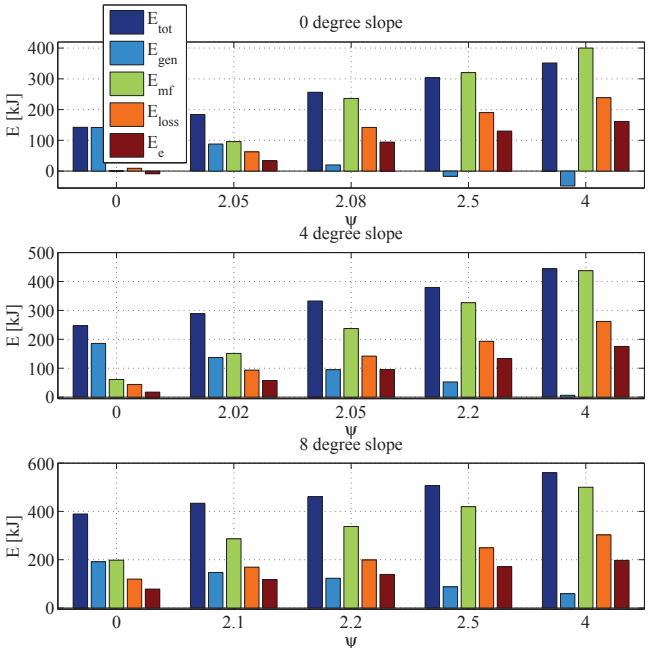


Figure 4. Energy balance for the candidate points in Figure 3.

4.4 Efficient state and control transients

The efficient state and control transients for the candidate points of the 0 degree slope case in Figure 3 are presented in Figure 5.

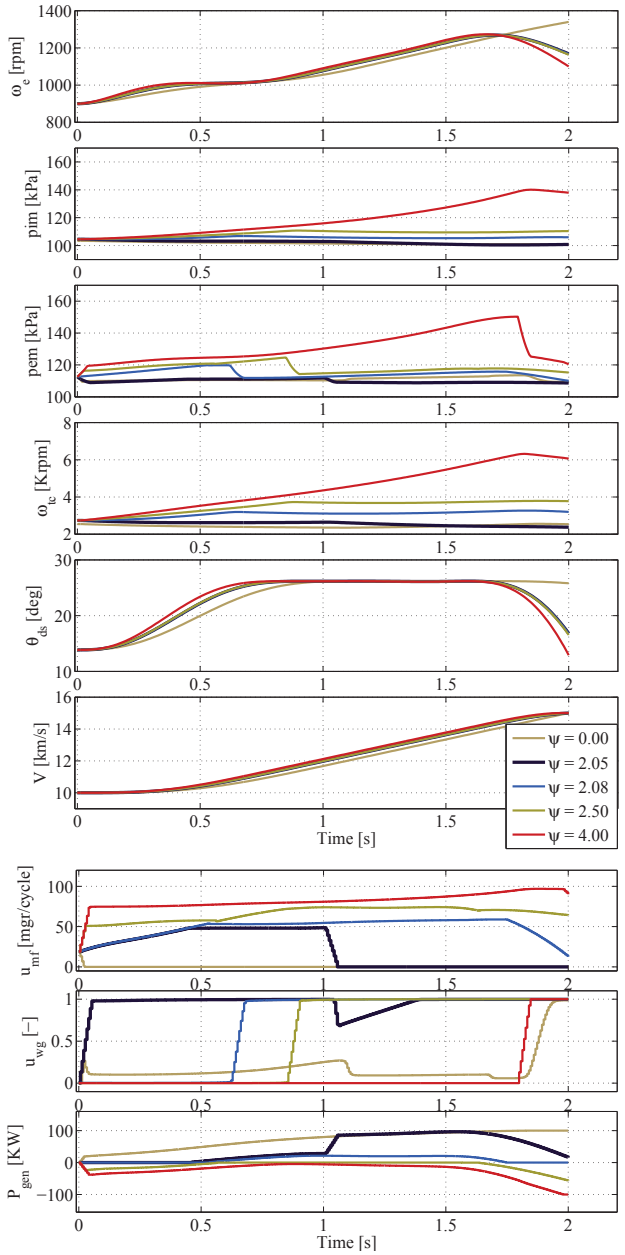


Figure 5. Optimal state and control transients during tip-in calculated for the candidate points in Figure 3.

According to the figures, as the cost for using electrical energy increases (larger ψ), the diesel engine transients are largely changed. for example comparing the $\psi = 0$ and $\psi = 4$ cases, when $\psi = 0$ fuel is cut-off, diesel engine power is not used and only electrical power accelerates the vehicle. But for $\psi = 4$, not only usage of the costly electrical energy is avoided but also parts of diesel engine power is used to store electrical energy at the end of the acceleration. Increasing the fuel conversion efficiency in the diesel engine operation is the main priority here and for that, fuel injection is selected such that the engine operates at the smoke limit delivering as large power as possible. This is similar to the discussion in (Sivertsson and Eriksson, 2015a) and (Sivertsson and Eriksson, 2015b) stating that the smoke limit dictates the solution during large parts of the transients. After an initial high power production which has facilitated fast vehicle acceleration, the wastegate which has been kept closed until this point, is opened at ca 1.7 [s] to lower the pumping work losses. Vehicle acceleration is reduced and less power from the engine is required to meet the final speed constraint. Instead, the engine power is used to build up electrical energy.

Considering the points mentioned about the extreme transients such as oscillatory controls or large acceleration at the end time, according to Figure 5, the transients for the $\psi \approx 2.05$ can be an example of improved control strategy with simple control transients and smooth vehicle speed transients at end time.

5 Conclusions

Optimal control of a diesel-electric powertrain during a tip-in acceleration is analyzed while importance of proper objective function formulation is highlighted. The extreme transients resulting from minimization of only jerk or time or energy are presented and the drawbacks in terms of oscillatory control signals are discussed. It is shown that by calculation of the trade-off between time and Jerk, low Jerk transients can be obtained. By applying penalties on energy consumptions in the Jerk minimization problem and solving for various fuel and electric energy weights in the objective function formulation, energy efficient transients are obtained. The calculated transients using this approach are presented which are simpler and more insightful for control design in real world applications. At the same time, the proposed controls maintain low Jerk and energy consumption compared to the extreme cases.

Acknowledgment

The support and feedback from engineers at SCANIA CV AB, and funding from Swedish Energy Agency is gratefully acknowledged.

References

Joel Andersson. *A General-Purpose Software Framework for Dynamic Optimization*. PhD thesis, Arenberg Doctoral School, KU Leuven, Department of Electrical Engineering (ESAT/SCD) and Optimization in Engineering Center, Kasteelpark Arenberg 10, 3001-Heverlee, Belgium, October 2013.

A Haj-Fraj and F Pfeiffer. Optimal control of gear shift operations in automatic transmissions. *Journal of the Franklin Institute*, 338(2):371–390, 2001.

A Haj-Fraj and F Pfeiffer. A model based approach for the optimisation of gearshifting in automatic transmissions. *International journal of vehicle design*, 28(1):171–188, 2002.

V. Nezhadali and L. Eriksson. Optimal control of engine controlled gearshift for a diesel-electric powertrain with backlash. In *AAC'16 – 8th IFAC Symposium on Advances in Automotive Control*, Kolmården, Sweden, 2016.

Magnus Pettersson and Lars Nielsen. Gear shifting by engine control. *IEEE Transactions Control Systems Technology*, 8 (3):495–507, May 2000.

Martin Sivertsson and Lars Eriksson. Time and fuel optimal power response of a diesel-electric powertrain. In *ECOSM'12 – IFAC Workshop on Engine and Powertrain Control, Simulation and Modeling*, Paris, France, October 2012a.

Martin Sivertsson and Lars Eriksson. Optimal step responses in diesel-electric systems. In *Mechatronics'12 – The 13th Mechatronics Forum International Conference*, Linz, Austria, September 2012b.

Martin Sivertsson and Lars Eriksson. Modeling for optimal control: A validated diesel-electric powertrain model. In *SIMS 2014 - 55th International Conference on Simulation and Modelling*, Aalborg, Denmark, 2014.

Martin Sivertsson and Lars Eriksson. Optimal transient control trajectories in diesel-electric systems-part 1: Modeling, problem formulation and engine properties. *Journal of Engineering for Gas Turbines and Power*, 137(2), February 2015a.

Martin Sivertsson and Lars Eriksson. Optimal transient control trajectories in diesel-electric systems-part 2: Generator and energy storage effects. *Journal of Engineering for Gas Turbines and Power*, 137(2), February 2015b.

Standardised On-Road Test cycles (SORT). `http://www.uitp.org/sites/default/files/documents/Knowledge/UITP_Project_SORT_GAS_20151021.pdf`, Last accessed April 2016.

Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1):25–57, 2006.

Numerical Efficiency of Inverse Simulation Methods applied to a Wheeled Rover

Thaleia Flessa Euan McGookin Douglas Thomson Kevin Worrall

Division of Aerospace Sciences, School of Engineering
University of Glasgow, UK, G12 8QQ

Abstract

A control method based on Inverse Simulation is applied to a four wheel rover. The method calculates the required inputs to achieve a desired, specified response; a trajectory in this case. Inverse Simulation considers the complete system dynamics to calculate the control input using an iterative, numerical Newton – Raphson scheme. Two methods for applying Inverse Simulation are presented, one based on a Differentiation scheme and one on Integration. The paper provides an insight into how the scheme formulation and selected parameters affect both methods' performance when applied to a rover. The selection of system outputs to control, their effect on each scheme's Jacobian, whether it is square or over-determined and the best method to factorize this Jacobian are investigated. The influence of the discretisation step and the convergence tolerance is also examined using two different sets for both schemes and in conjunction with the type of Jacobian used. The comparison is made in terms of the resulting trajectory, the execution time, and the quality of the calculated control input.

Keywords: inverse, simulation, control, navigation, model based, numerical, wheeled vehicle, rover

1 Introduction

A novel method based on Inverse Simulation is used for planetary rover guidance and control. Inverse simulation uses a mathematical model that is representative of the system and calculates the control inputs necessary to produce the desired response. This desired response is defined in terms of the system's output variables and represents their time history. Inverse Simulation is a model based, numerical, iterative process where step changes in the various controls are applied until the predicted response matches the desired response (Thomson and Bradley, 2006). Applied to rover navigation, the desired response is a specified, safe trajectory to a goal destination (Worrall *et al.*, 2015a; Worrall *et al.*, 2015b). Inverse Simulation is a novel way of addressing the issue of given a specified, safe path, what are the required control inputs for the rover to reach the destination goal through this path. The method can be applied (a) in situ: given a series of waypoints or a defined trajectory, the rover can calculate the necessary control inputs or (b) offline: operators define the trajectory, the control inputs are calculated and then sent to the rover.

Applications for Inverse Simulation are predominantly within the flight dynamics domain and the application to rotorcraft flight control is a major area. In these particular cases Inverse Simulation is used to produce the required control signals for specific flight manoeuvres (Hess and Gao, 1993; Murray-Smith, 2000; Thomson and Bradley, 2006) and (Avanzini *et al.*, 2013) also introduces a predictive element. The method has also been applied to unmanned aerial vehicles (Murray-Smith and McGookin, 2015) and autonomous underwater vehicles (Murray-Smith *et al.*, 2008). Inverse Simulation has also been used as a model validation method (Murray-Smith, 2000; Thomson and Bradley, 2006). Previous research has demonstrated the potential for Inverse Simulation as a guidance and control method for wheeled rovers (Worrall *et al.*, 2015a; Worrall *et al.*, 2015b).

Planetary rover navigation so far has been achieved using a combination of non-, semi- and fully autonomous methods (Bajracharya *et al.*, 2008). The NASA Mars Exploration Rovers (MER) use a combination of three main driving modes with varying degrees of autonomy. The first mode involves the rover executing a sequence of commands to follow a defined course of waypoints towards specific goal coordinates. In this mode the rover only performs basic safety checks (Biesiadecki *et al.*, 2007). The second mode is semi-autonomous navigation during which the rover is given a set of waypoints towards specific goal coordinates and uses its on-board capabilities for hazard avoidance and for planning a path towards the goal. A special case is when the rover drives towards an area that is unknown to the operators (Biesiadecki *et al.*, 2007; Bajracharya *et al.*, 2008). In this case the rover has to choose the waypoints for a safe path towards the goal and then drive along this path; this is fully autonomous navigation. The third mode is visual odometry: the rover uses images from the on-board cameras to accurately estimate and update its position

(Cheng *et al.*, 2005; Biesiadecki *et al.*, 2007). A similar combination of these driving modes is used for the Curiosity rover and autonomous navigation is used to plot a safe path towards an area unknown to the operators (Bakambu *et al.*, 2012). The fully autonomous and visual odometry modes are used when the rover moves into areas that are not visible to the operators (Cheng *et al.*, 2005; Biesiadecki *et al.*, 2007; Bajracharya *et al.*, 2008). The developers of the ExoMars mission have addressed the issue of control, navigation and autonomy by including an element of autonomous control (Silva *et al.*, 2013) and by conducting field experiments (Woods *et al.*, 2014). Another issue is the computational complexity of the algorithms that are running on-board. The MER and Curiosity rovers use special space qualified and radiation hardened microprocessors whose computational capabilities have been exceeded by more than two orders of magnitude by the average desktop computer (Howard *et al.*, 2012). Furthermore, the navigation algorithms must also be tested using a wide range of parameters, which is best done using simulation (Madison *et al.*, 2007).

The paper investigates the selection of outputs to control and the parameters that affect the application and execution time of Inverse Simulation to a four wheeled rover. The control inputs are calculated from Inverse Simulation, applied to the rover and the resulting trajectory is compared with the desired.

2 Methodology of Inverse Simulation

Inverse Simulation has two main requirements: a desired trajectory represented as a time history with an appropriate time step and a model of the system. The model's inputs and outputs must be representative of the inputs and outputs of the actual system. The desired trajectory is described using the model's outputs. There are two main implementations of Inverse Simulation for finding the control inputs given a desired output: Differentiation (Hess and Gao, 1993; Murray-Smith, 2000; Thomson and Bradley, 2006; Murray-Smith and McGookin, 2015) and Integration (Hess and Gao, 1993; Thomson and Bradley, 2006; Avanzini *et al.*, 2013; Worrall *et al.*, 2015a; Worrall *et al.*, 2015b). The basic framework for each is similar and uses a numerical Newton – Raphson algorithm; what differs is the method of convergence to the control signal. In Differentiation, a numerical differentiation scheme is used and the convergence is based on the system's state and output equations. In Integration, a numerical integration scheme is used and the convergence is based on whether the system's output matches the desired. An alternative approach to Inverse Simulation uses a modified version of the Integration scheme and search based optimisation (Lu *et al.*, 2008).

2.1 Implementation

A general non-linear system is used where $\boldsymbol{f}\in\mathbf{R}^{\mathbf{m}}$ are the state equations, $\boldsymbol{g}\in\mathbf{R}^{\mathbf{p}}$ are the output equations, $\boldsymbol{u}\in\mathbf{R}^{\mathbf{q}}$ is the control input vector, $\boldsymbol{x}\in\mathbf{R}^{\mathbf{m}}$ is the state variable vector and $\boldsymbol{y}\in\mathbf{R}^{\mathbf{p}}$ is the output vector. The desired output to control is $\boldsymbol{g_d}\in\mathbf{R}^{\mathbf{p}}$.

$$\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x},\mathbf{u}),\ \mathbf{y}=\mathbf{g}(\mathbf{x},\mathbf{u}) \tag{1}$$

For the Differentiation method, (1) is discretised N times over a time interval T, where dt is the discretisation step. The unknowns in (2) are the states $\boldsymbol{x}$ and the input $\boldsymbol{u}$ at t_i. The known variables are the desired output $\boldsymbol{g_d}$ and the states, control and output from the previous discretisation step t_{i-1}. Then, the functions $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ in (3) are defined to find the values of input $\boldsymbol{u}$ and the states $\boldsymbol{x}$ for the given output $\boldsymbol{g_d}$. The system in (3) is solved using the Newton - Raphson method to update $\boldsymbol{u}$ and $\boldsymbol{x}$ until their values are such that $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ are both equal to zero within a certain tolerance. The updated equations are in (4) and $\boldsymbol{J}$ is the Jacobian of the system from (3).

$$\begin{aligned}&\frac{\mathbf{x}(t_i)-\mathbf{x}(t_{i-1})}{dt}=\mathbf{f}\left(\mathbf{x}(t_i),\mathbf{u}(t_i)\right),\ dt=t_i-t_{i-1}\\&\mathbf{y}(t_i)=\mathbf{g}\left(\mathbf{x}(t_i),\mathbf{u}(t_i)\right)\end{aligned} \tag{2}$$

$$\begin{aligned}&\mathbf{F}_1=\mathbf{f}\left(\mathbf{x}(t_i),\mathbf{u}(t_i)\right)-\frac{\mathbf{x}(t_i)-\mathbf{x}(t_{i-1})}{dt}\\&\mathbf{F}_2=\mathbf{g}\left(\mathbf{x}(t_i),\mathbf{u}(t_i)\right)-\mathbf{g}_d(t_i)\end{aligned} \tag{3}$$

$$\begin{bmatrix}\mathbf{x}_n\\\mathbf{u}_n\end{bmatrix}(t_i)=\begin{bmatrix}\mathbf{x}_{n-1}\\\mathbf{u}_{n-1}\end{bmatrix}-\mathbf{J}^{-1}\cdot\begin{bmatrix}\mathbf{F}_1(\mathbf{x}_{n-1},\mathbf{u}_{n-1})\\\mathbf{F}_2(\mathbf{x}_{n-1},\mathbf{u}_{n-1})\end{bmatrix}(t_i) \tag{4}$$

For the Integration approach the state and output equations from (1) are again discretised and dt is the discretisation step. The state equations are integrated at t_i. An error function between the current output and the desired $\boldsymbol{g_d}$ is defined in (6). Equation (6) is solved for $\boldsymbol{u}$ using the Newton – Raphson method and the iterative relationship in (7). $\boldsymbol{J_e}$ is the Jacobian of the error function $\boldsymbol{f_e}$ or equivalently the Jacobian of the system outputs when perturbing the inputs.

$$\begin{aligned}&\mathbf{x}(t_i)=\int_{t_{i-1}}^{t_i}\dot{\mathbf{x}}(\tau_i)\,d\tau+\mathbf{x}(t_{i-1})\\&\mathbf{y}(t_i)=\mathbf{g}\left(\mathbf{x}(t_i),\mathbf{u}(t_{i-1})\right)\end{aligned} \tag{5}$$

$$\mathbf{f_e}=\mathbf{g}\left(\mathbf{x}(t_i),\mathbf{u}(t_{i-1})\right)-\mathbf{g_d}(t_i) \tag{6}$$

$$\mathbf{u}_n(t_{i-1})=\mathbf{u}_{n-1}-\mathbf{J}_e^{-1}(\mathbf{x}_{n-1},\mathbf{u}_{n-1})\cdot\mathbf{f_e}(\mathbf{x}_{n-1},\mathbf{u}_{n-1}) \tag{7}$$

2.2 Numerical Properties

Both implementations use a Jacobian and care must be taken when trying to find its inverse or a suitable

factorization. For the Differentiation method, from Eq. (4) the dimension of the Jacobian $\boldsymbol{J}$ is $[m+p]\times[m+q]$. For the Integration method from Eq. (7) the dimension of the Jacobian $\boldsymbol{J_e}$ is $[p]\times[q]$. If there is an equal number of inputs and outputs ($p=q$), then the Jacobian is a square matrix. If however, the number of inputs and outputs is not equal, then factorization methodologies such as LU, QR or Cholesky decomposition or the Moore–Penrose pseudoinverse (Strang, 2009; Davis, 2013), can be used. When there are more outputs than inputs ($p>q$), this results in an over-determined system and the pseudoinverse or factorization can be used. In that case, the calculated outputs are a least-square fit to the desired outputs and not necessarily a good one. For this reason, systems where the number of inputs is equal to or greater than the number of outputs are preferred candidates (Hess and Gao, 1993; Murray-Smith, 2000; Thomson and Bradley, 2006; Murray-Smith *et al.*, 2008).

Each approach has advantages and disadvantages, which are usually identified as the following (Hess and Gao, 1993; Murray-Smith, 2000; Thomson and Bradley, 2006; Lu *et al.*, 2008): (a) The Integration method can use any representative model of the system. Differentiation requires both the states and the outputs and any change in the model results in a reformulation of the algorithm. Therefore, the Differentiation method is more time consuming to set up and maintain, whereas for Integration the model can be modified more easily, (b) The Integration method has a convergence rate that is up to an order of magnitude larger than that of the Differentiation method but it is generally more stable; what is gained in flexibility and stability, is lost in computing time. The numerical properties of Inverse Simulation have been examined mostly when the method is applied to flight dynamics (Hess and Gao, 1993; Thomson and Bradley, 2006; Lu *et al.*, 2008). It was observed that there are oscillations in the response of the uncontrolled states (constraint oscillations) (Thomson and Bradley, 2006; Lu *et al.*, 2008). However, these oscillations depend more on the dynamical properties of the system, its uncontrollable states and zero dynamics as well as on the discretisation step *dt*, rather than on the method used. They are also significantly reduced when a larger *dt* is used (Lu *et al.*, 2008). Also from (Thomson and Bradley, 2006) it was observed that there are low amplitude, high frequency oscillations superimposed on the calculated control input. These oscillations are due to several reasons (Hess and Gao, 1993; Murray-Smith, 2000; Thomson and Bradley, 2006; Lu *et al.*, 2008): redundancy issues, non-square Jacobian and multiple solutions, several local minima of the error function from (4), (7). The oscillations increase when the discretisation step *dt* is too small, as it could excite the uncontrollable states (Lu *et al.*, 2008). Nonetheless, a relatively small *dt* can have a positive effect because it captures the changes in the system dynamics and this may reduce or even remove them (Lu *et al.*, 2008).

3 Rover Model and Trajectory Generation

Inverse Simulation requires a mathematical model of the system and a desired response, which is a trajectory. First, a path to the destination is determined as a series of waypoints. This information provides the desired trajectory for the Inverse Simulation, which in turn generates the required guidance commands (control inputs) to follow the trajectory (Worrall *et al.*, 2015a; Worrall *et al.*, 2015b).

3.1 Rover Model

The model of the rover has been presented in (Worrall, 2010; Worrall *et al.*, 2015a; Worrall *et al.*, 2015b) and has been experimentally validated (Worrall, 2010). It is briefly described here for clarity. Each side has two wheels and the wheels at each side provide the same torque input. The dynamics are described by (8), where $\boldsymbol{v}$ is the state velocity vector (9) in the local body frame, $\boldsymbol{\eta}$ is the velocity vector in the global frame and $\boldsymbol{\tau}$ is the input vector (10).

$$\begin{bmatrix}\dot{\mathbf{v}}\\ \dot{\boldsymbol{\eta}}\end{bmatrix}=\begin{bmatrix}\mathbf{M}^{-1}\{\boldsymbol{\tau}-\mathbf{C}(\mathbf{v})\mathbf{v}-\mathbf{D}(\mathbf{v})\mathbf{v}-\mathbf{g}(\boldsymbol{\eta})\}\\ \mathbf{J}_t(\boldsymbol{\eta})\mathbf{v}\end{bmatrix} \quad (8)$$

$$\mathbf{v}=\begin{bmatrix}u & v & w & p & q & r\end{bmatrix}^T \quad (9)$$

$$\boldsymbol{\tau}=\begin{bmatrix}X & Y & Z & K & M & N\end{bmatrix}^T \quad (10)$$

In (9) u, v, w are the surge, sway and heave velocities and p, q, r are the roll, pitch and yaw rates respectively. In (10) X is the surge, Y is the sway and Z is the heave force, K is the roll, M is the pitch and N is the yaw moment. X and N are controllable by two inputs: one torque at each side. The remaining forces and moments are the unmatched dynamics.

3.2 Trajectory Generation

The trajectory is represented as a series of waypoints, each defined by an x-y coordinate with a common origin. A path between each waypoint and the next is calculated, with the robot stopping at each waypoint to turn on the spot to achieve the desired orientation and then move again.

The distance and time to travel between each waypoint is calculated assuming a constant velocity between stages with initial and final acceleration and deceleration transients: the constant forward speed is 0.1ms^{-1}, analogous to that of operating rovers, and the rotational velocity is 0.1rads^{-1}. At each waypoint a check is made to determine if the rover is at the correct angle for the next traversal forward. If not, then the

rover is commanded to turn on the spot until the desired angle is achieved. The path from one waypoint to the next is defined by specifying the acceleration as a 7th order polynomial function of time and is based on that presented in (Thomson and Bradley, 2006; Worrall *et al.*, 2015a). A 7th order polynomial has the benefit of producing smooth trajectory profiles with high order, continuous derivatives. The output is the acceleration time history, which is then integrated to provide the velocities and the displacements. The result is a continuous time history of acceleration, speed and distance between each successive waypoint that fully describes the rover's position and orientation; namely the elements of $\boldsymbol{v}$, $\boldsymbol{\eta}$.

4 Application Results

A series of waypoints are first defined and then a trajectory between them is generated as in Section 3. It is assumed that these waypoints represent a safe, feasible path. Inverse Simulation calculates the control inputs for each trajectory. Then these inputs are applied to the system and it is checked whether the resulting trajectory matches the desired. The following test trajectories were selected. The Long Distance test (Figure 1, 400s) involves several pose changes and will be used as a benchmark to show how the errors built up over time and to compare the different parameters. The Figure of Eight test (Figure 2, 175s) is used to demonstrate a complex path with multiple, sharp turns and how the method copes with successive pose changes.

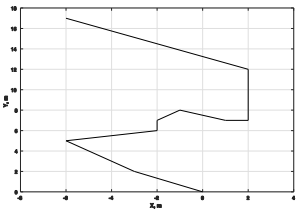


Figure 1. Long Distance.

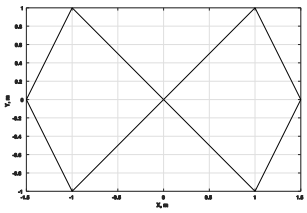


Figure 2. Figure of Eight.

The simulation parameters that need to be assigned values are: discretisation step *dt*, convergence tolerance *tol*, torque input initial estimate, maximum number of iterations for the Newton-Raphson algorithm. For *dt*, the timestep of the motors and the need to adequately follow the system are taken into account (Worrall *et al.*, 2015a). The rover starts from rest (zero motor torque). Here a very small value is set for the initial estimate. The number of iterations is set to ensure convergence without increasing the execution time.

The assessment criteria are the following: (a) mean error and standard deviation between the actual and the desired position x (integrated from u), (b) mean error and standard deviation between the actual and the desired heading angle θ (integrated from r), (c) calculation time, an important measurement for any control algorithm. The position and heading angle represent the rover position in space and hence how wells it follows the desired trajectory.

Table 1. Simulation Parameters.

Parameter	*Set 1*	*Set 2*
dt (s)	0.01	0.05
tol	5×10^{-7}	5×10^{-5}
initial control estimate (Nm)	2.5×10^{-7}	
maximum iterations	30	
MATLAB version	2014b, 64 bit	
hardware	Core 2 Duo T9300, 2.50 GHz, 4 GB RAM	

4.1 Selection of inputs, outputs and Jacobian inversion.

The two controllable outputs are the surge X and the yaw moment N. This is equivalent to controlling the surge and yaw velocities and so the desired outputs are u_d, r_d. There are two control inputs, one torque per side (τ_{left}, τ_{right}). For the Integration method, there are two inputs and two outputs and so the size of the Jacobian (Worrall *et al.*, 2015a; Worrall *et al.*, 2015b) in Eq. (7) is 2x2. For the Differentiation method, it was observed during the initial simulations that including as an additional output to control the sway velocity v, the overall results are significantly improved without sacrificing greatly in execution time. There are three desired outputs: u_d, r_d as before and v_d, which is set to zero. The sway velocity v is not matched dynamically to the actuators of the system and therefore cannot be directly controlled. It is however strongly coupled to u and r (Worrall, 2010) and this interaction can provide indirect control of sway and act as an additional constraint. For the Jacobian, Eq. (4), only the controllable states u, r and also v are taken into account and so its size is 6x5. The remaining states for (4) are estimated after convergence at each t_i. This is an over-determined system and to ensure that the solution is always a least square solution a suitable factorization method is used to find the pseudo-inverse of $\boldsymbol{J}$ and solve (4). There are several methods to find the Jacobian inverse. Table 2 shows the methods in MATLAB that are examined (Davis, 2013). Each method from Table 2 is tested using the Long Distance test and the first set of parameters.

Table 3 shows the results for the Differentiation scheme. The *backslash* method fails because $\boldsymbol{J}$ is (column) rank deficient; this is expected because the outputs to control are u, v and r and v is strongly coupled to u and r. Between *pinv(**J**)* and *factorize(**J**)*, the factorize command is superior in terms of errors and is the one selected, at the expense of increased execution time. The method used by *factorize(**J**)* is the complete orthogonal decomposition, which is suitable for rank deficient systems. Table 4 shows the results for

the Integration scheme. The *backslash* method is the best for the error and execution time and is the one selected. Integration is slower, which is in line with previous observations (Section 2.2).

Table 2. Inversion Methods.

Method	*Comments*
inv() - built-in function	Suitable only for square systems of full rank, can be very inaccurate.
pinv() - built-in function	Suitable for non-square systems, calculates the Moore–Penrose pseudoinverse using singular value decomposition (SVD).
\ (backslash operator) - built-in function	Suitable for square or over determined systems with full column rank, fast, accurate. Factorization cannot be reused. Suggested MATLAB method.
FACTORIZE (Davis, 2013) - additional package, acts as a wrapper for the built-in MATLAB functions	Selects the most suitable factorization method from LU decomposition, Cholesky decomposition, QR decomposition, SVD (singular value decomposition), COD (complete orthogonal decomposition). Suitable for square, rank deficient and over/under determined systems.

Table 3. Inversion: Differentiation (with sway), Long Distance (set 1).

	\	*pinv(J)*	*factorize(J)*
mean position x error (m)	-	0.00247	0.00082
σ position error	-	0.00211	0.00080
mean heading θ error (rad)	-	0.00123	0.00053
σ heading error	-	0.00076	0.00056
execution time (s)	-	34.12	55.45

Table 4. Inversion: Integration (without sway), Long Distance (set 1).

	\	*inv(J)*	*factorize(J)*
mean position x error (m)	0.00082	0.00082	0.00082
σ position error	0.00040	0.00040	0.00040
mean heading θ error (rad)	0.00073	0.00073	0.00073
σ heading error	0.0013	0.0013	0.0013
execution time (s)	79.19	376.89	317.84

4.2 Scheme comparison

From Tables 3, 4 the main difference between the two schemes is the calculation time. Differentiation performs slightly better for the heading. Figure 3 shows the control inputs generated for Differentiation and Figure 4 for Integration.

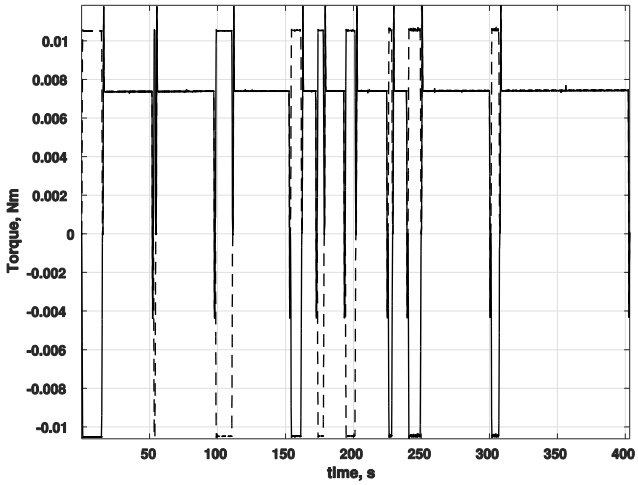


Figure 3. Differentiation Control, Long Distance (set 1).

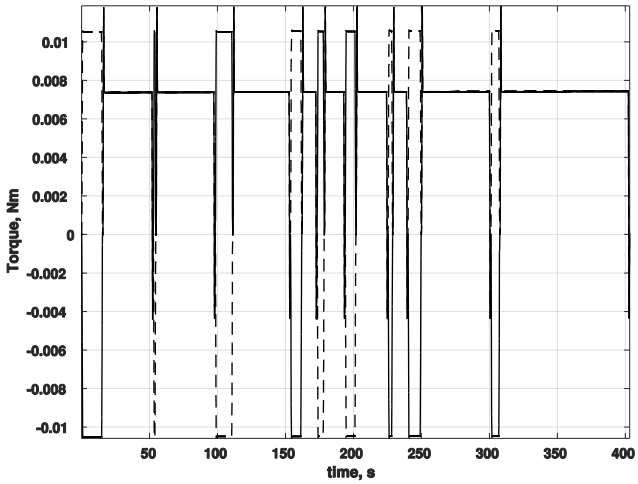


Figure 4. Integration Control, Long Distance (set 1).

The left side control is signified by the solid line and the right side by the dashed line. The left and right signals are symmetrical when the rover is moving forward (e.g. at 100s), which is expected since each side is controlled by one input. When the heading changes there is a momentary spike in the input. The control inputs from Integration are smoother, e.g. between 100 – 150s and at 250s in Figure 3 and 4. The oscillations from Differentiation have a small magnitude and high frequency and are due to the fact that the scheme uses an over-determined system which may have multiple solutions (Section 2.2).

Table 5 shows the results for the Long Distance test, set 2: *dt* is increased and so is the convergence tolerance. The execution time is significantly reduced, and Integration is now faster than Differentiation. Integration performs slightly better in terms of the

position error and Differentiation is better for the heading error. Compared with Tables 3, 4, the standard deviation of both the position and the heading error is larger; the rover has some sharper deviations from the desired position and heading. Figure 5 shows the calculated control input from Differentiation. By increasing the *dt* to 0.05s, the high frequency, low amplitude oscillations in the control input decrease significantly.

Table 5. Long Distance (set 2).

	Differentiation (with sway)	*Integration (without sway)*
mean position x error (m)	0.00468	0.00450
σ position error	0.00314	0.00196
mean heading θ error (rad)	0.00276	0.00736
σ heading error	0.00262	0.00700
execution time (s)	12.19	8.72

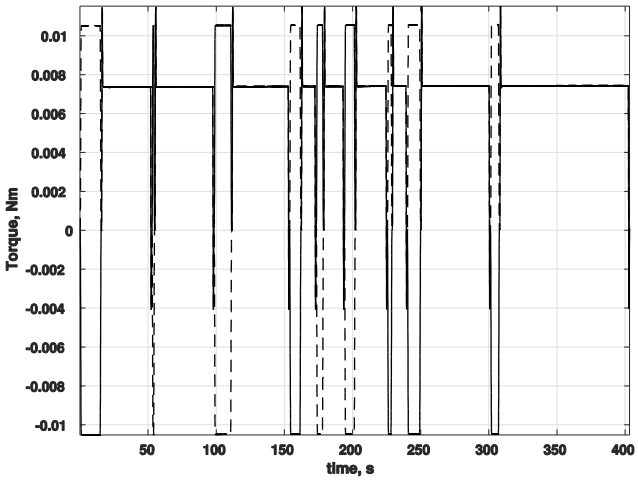


Figure 5. Differentiation Control, Long Distance (set 2).

Table 6 shows the results for the Figure of Eight test. Both methods perform similarly but Differentiation is faster and slightly better for the standard deviation of the heading error.

Figure 6 and 7 show the errors of *u*, *v* and *r* after the Newton–Raphson algorithm for Eq. (3) and (6) has converged at each t_i. The desired value of *v* is set to zero and the desired values of *u* and *r* are the same for both methods. For Differentiation, the error in *r* deviates about 10^{-16} rad/s from zero, whereas for Integration it deviates about 10^{-4} rad/s from zero. The *v* error deviates 10^{-7} m/s from zero and the *u* error deviates 10^{-8} m/s from zero for Differentiation. For Integration, the *v* error deviates 10^{-4} m/s and the *u* error 10^{-5} m/s from zero. At Figure 7, when the heading changes there is a spike in the errors. At Figure 6, the much smaller errors are due to *v* used as an additional output. This effect is particularly evident when comparing the errors in *r*.

Table 6. Figure of Eight (set 1).

	Differentiation (with sway)	*Integration (without sway)*
mean position x error (m)	0.00034	0.00060
σ position error	0.00036	0.00048
mean heading θ error (rad)	0.00202	0.00203
σ heading error	0.09501	0.11632
execution time (s)	25.94	40.45

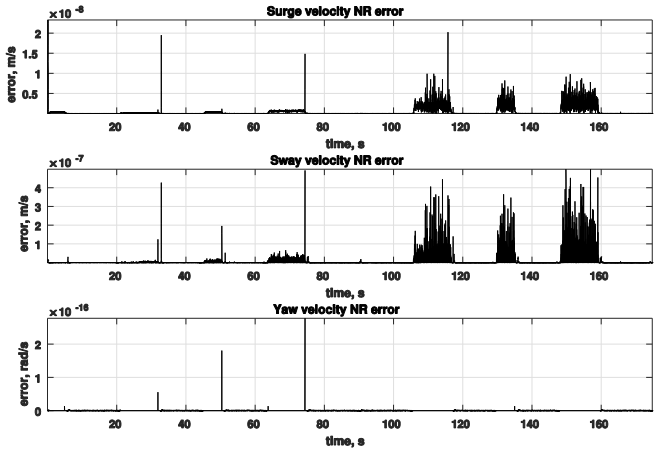


Figure 6. Differentiation NR Errors Fig of Eight (set 1).

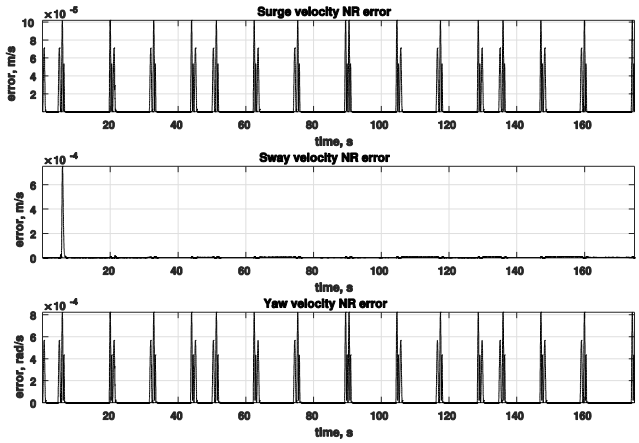


Figure 7. Integration NR Errors Fig. of Eight (set 1).

From Tables 3 - 6, Integration exhibits bigger errors for the heading angle. When the *dt* and the tolerance are small enough or when there are no abrupt orientation changes, this is negligible. As *dt* and the tolerance increase and the trajectory requires sharp heading changes, this difference becomes more important. This can be seen in the failure of Integration for the Figure of Eight test using parameter set 2. The errors in *r* are not corrected, the calculated control in (7) increases and the condition number of $\boldsymbol{J_e}$ by 22.5s (total time 175s) is infinite; the Jacobian is ill-conditioned and cannot be factorized.

4.3 Effect of sway velocity

To reduce the error in r and conversely in θ, the sway velocity v is used as an additional output for Integration. Then, $\boldsymbol{J_e}$ in Eq. (7) has a size of 3x2: three outputs (u, v, r), two inputs and the *factorize* command is used. Table 7 shows the results for the Figure of Eight test, set 2. Both methods produce similar results, however Integration is slower and still has a larger error and standard deviation for the heading. Nonetheless, the usage of v has here a positive effect and enables Integration to converge. When using v in Integration for the Long Distance test, set 2, the execution time increases to 29.71s compared to 8.72s (Table 5) without any error improvements. For the Long Distance test, set 1 (Table 4), the time is greatly increased to 791.99s.

Table 7. Figure of Eight Test (set 2).

	Differentiation (with sway)	*Integration (with sway)*
mean position x error (m)	0.00384	0.00332
σ position error	0.00259	0.00231
mean heading θ error (rad)	0.01047	0.02298
σ heading error	0.00259	0.27941
execution time (s)	5.30	20.71

Table 8. Differentiation: Long Distance (set 1).

	Differentiation (with sway)	*Differentiation (without sway)*
mean position x error (m)	0.00082s	0.002404
σ position error	0.00080	0.001960
mean heading θ error (rad)	0.00053	0.003154
σ heading error	0.00056	0.002864
execution time (s)	55.45	56.26

Table 8 shows the Differentiation results for the Long Distance test (set 1) with and without using v. Without v, $\boldsymbol{J}$ in (4) is square (4x4). The errors increase by two orders of magnitude and the execution time is almost the same: the method converges slower and with larger errors. Compared with Integration (Table 4), the inclusion of v has a greater effect on Differentiation. This confirms previous results, that Integration is more stable. Here, Differentiation performs slightly better but requires an over-determined system and specialized handling.

5 Conclusions

The selection of outputs to control, their effect on the size of the Jacobian and the best factorization method were examined. A square Jacobian is used for Integration and an over-determined for Differentiation. The schemes were compared for varying dt and convergence tolerance. A small dt results in high frequency, low amplitude oscillations in the control input from Differentiation. To remove these, the dt was increased. The effect of sway velocity v, which is strongly coupled with u, r but not directly controlled, was examined. For Differentiation, using v as an output is beneficial from the start. Integration performed well for both parameter sets for the Long Distance test without v. For the Figure of Eight test for a dt of 0.05s and tolerance 5×10^{-5}, including v was necessary. It is worth noting that this test is not a realistic trajectory and is used to test the method's limits. A dt of 0.01s and a tolerance of 5×10^{-7} produce the best results, with increased calculation time. For simplicity and overall stability, the Integration scheme is more appropriate. For decreased execution time, Differentiation is preferred, at the expense of slightly larger position errors and an over-determined system. In all cases, the control inputs from Inverse Simulation where within operational limits.

Acknowledgments

Research supported by grant EPSRC/1369575 from the UK Engineering and Physical Sciences Research Council (EPSRC).

References

G. Avanzini, D. G. Thomson, and A. Torasso. Model Predictive Control Architecture for Rotorcraft Inverse Simulation. *Journal of Guidance, Control, and Dynamics*, 36(1), 207–217, 2013. doi:10.2514/1.56563.

M. Bajracharya, M. W. Maimone, and D. Helmick. Autonomy for Mars Rovers: Past, Present, and Future. *Computer*, 41(12), 44–50, 2008. doi: 10.1109/MC.2008.479.

J. N. Bakambu, C. Langley, G. Pushpanathan, W. J. MacLean, and R. Mukherji. Field trial results of planetary rover visual motion estimation in Mars analogue terrain. *Journal of Field Robotics*, 29(3), 413–425, 2012. doi: 10.1002/rob.21409.

J. J. Biesiadecki, P. C. Leger, and M. W. Maimone. Tradeoffs between Directed and Autonomous Driving on the Mars Exploration Rovers. *The International Journal of Robotics Research*, 26(1), 91–104, 2007. doi: 10.1177/0278364907073777.

Y. Cheng, M. W. Maimone, and L. Matthies. Visual Odometry on the Mars Exploration Rovers. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, Waikoloa, HI, USA, pages 903–910, 2005. doi: 10.1109/ICSMC.2005.1571261.

T. A. Davis. Algorithm 930: FACTORIZE: An Object-Oriented Linear System Solver for MATLAB. *ACM Transactions on Mathematical Software*, 39(4), 1–18, 2013. doi: 2491491.2491498.

R. A. Hess and C. Gao. A Generalized Algorithm for Inverse Simulation Applied to Helicopter Maneuvering Flight. *Journal of the American Helicopter Society*, 38(4), 3–15, 1993. doi: 10.2514/3.20732.

T. M. Howard, A. Morfopoulos, J. Morrison, Y. Kuwata, C. Villalpando, L. Matthies, and M. McHenry. Enabling continuous planetary rover navigation through FPGA stereo and visual odometry. In *2012 IEEE Aerospace Conference*, Big Sky, MT, USA, pages 1–9, 2012. doi: 10.1109/AERO.2012.6187041.

L. Lu, D. J. Murray-Smith, and D. G. Thomson. Issues of numerical accuracy and stability in inverse simulation. *Simulation Modelling Practice and Theory*, 16(9), 1350–1364, 2008. doi: 10.1016/j.simpat.2008.07.003.

R. Madison, A. Jain, C. Lim, and M. W. Maimone. Performance characterization of a rover navigation algorithm using large-scale simulation. *Scientific Programming*, 15(2), 95–105, 2007. doi: 10.1155/2007/638280.

D. J. Murray-Smith. The inverse simulation approach: a focused review of methods and applications. *Mathematics and Computers in Simulation*, 53(4–6), 239–247, 2000. doi: 10.1016/S0378-4754(00)00210-X.

D. J. Murray-Smith, L. Lu, and E. W. McGookin. Applications of inverse simulation to a nonlinear model of an underwater vehicle. In *Summer Simulation Multi-Conference 2008 - Grand Challenges in Modelling and Simulation*, Edinburgh, Scotland, 2008.

D. J. Murray-Smith and E. W. McGookin. A case study involving continuous system methods of inverse simulation for an unmanned aerial vehicle application. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 229(14), 2700–2717, 2015. doi: 10.1177/0954410015586842.

N. Silva, R. Lancaster, and J. Clemmet. ExoMars Rover Vehicle Mobility Functional Architecture and Key Design Drivers. In *12th Symposium on Advanced Space Technologies in Robotics and Automation (ASTRA)*, Noordwijk, The Netherlands, 2013.

G. Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley MA, 4th ed. 2009.

D. G. Thomson and R. Bradley. Inverse simulation as a tool for flight dynamics research - Principles and applications. *Progress in Aerospace Sciences*, 42(3), 174–210, 2006. doi: 10.1016/j.paerosci.2006.07.002.

M. Woods, E. Tidey, B. Van Pham, L. Simon, R. Mukherji, B. Maddison, G. Cross, A. Kisdi, W. Tubby, G. Visentin, and G. Chong. Seeker-Autonomous Long-range Rover Navigation for Remote Exploration. *Journal of Field Robotics*, 31(6), 940–968, 2014. doi: 10.1002/rob.21528.

K. J. Worrall, D. G. Thomson, E. W. McGookin, and T. Flessa. Autonomous Planetary Rover Control using Inverse Simulation. In *13th Symposium on Advanced Space Technologies in Robotics and Automation (ASTRA 2015)*, Noordwijk, The Netherlands, 2015a.

K. J. Worrall. *Guidance and search algorithms for mobile robots: application and analysis within the context of urban search and rescue*. PhD Thesis, University of Glasgow.

K. J. Worrall, D. G. Thomson, and E. W. McGookin. Application of Inverse Simulation to a wheeled mobile robot. In *2015 6th International Conference on Automation, Robotics and Applications (ICARA)*, Queenstown, New Zealand, pages 155–160, 2015b. doi: 10.1109/ICARA.2015.7081140.

An Improved Kriging Model based on Differential Evolution

Xiaobing Shang, Ping Ma*, Ming Yang

Control and Simulation Center, Harbin Institute of Technology, P.R. China, shangxiaobing1992@126.com, pingma@hit.edu.cn, myang@hit.edu.cn

Abstract

Kriging model is a commonly interpolate approximation method which is widely used in the computer simulation in the past decade. The fitting accuracy is one of the fundamental problems in the research of kriging model, which can be summarized in two aspects, the accurate estimation of model's parameters and approximate form selection of kriging model. In order to solve the existed problems, an improved parameter estimation method of kriging model base on differential evolution (DE) algorithm is set out in the present paper. Firstly, establish the objective function of DE algorithm depends on the estimation of the model's accuracy, and get the optimum solution of model's parameters under the initial condition. Then, a variety of regression function and correlation function in kriging models are selected to compare the fitting accuracy. Finally, the simulation case for outer ballistic data on electromagnetic railgun is examined to determine whether the improved method has priority over traditional one in the approximation accuracy.

Keywords: Kriging model, DE algorithm, approximation accuracy, EM railgun

1 Introduction

Kriging interpolation model, which is an optimal linear unbiased estimate model, is a commonly spatial interpolation model based on the geostatistical variogram theory. In 1989, this theory is introduced to the computer simulation field, and then gradually become a frequently-used interpolation method, which is used in a variety of applications including mechanics engineering, structural optimization and sequential experimental design (Volpi *et al*, 2015). The approximation accuracy is one of the key problems of current research which including how to effectively estimate the kriging model's parameters and select the approximate form of model. These two aspects determine the fitting accuracy, which is the significant approach of model investigation.

A mounts of research is taken to solve the problems of improving the model's accuracy. Jay D. Martin proposed an estimate method of model's parameters by using maximum likelihood estimate (MLE) and cross validation (CV) methods. The analysis results showed that the MLE is prior to CV method, which are applied in three different dimensional fitting problems (Martin *et al*, 2005; Martin *et al*, 2004). Jack P.C. summarized currently research achievements and pointed out that the approximation accuracy is determined by the form and parameters of correlation function (Kleijnen *et al*, 2009). Søren N. Lophaven developed a matlab tool DACE, which used to compute the kriging model effectively. The model's parameters are made easy through DACE. Unfortunately, this tool need to limit the minimum and maximum value of parameters, and cannot find the optimal results, which determine the narrow application fields of this method (Lophaven *et al*, 2002; Lophaven *et al*, 2002). LIU Xiaolu considers the model's accuracy is not only determined by parameters estimation, but also by the sampling points. The improved general pattern search (IGPS) algorithm was used to get these points. And a satellite orbit parameter optimization problem is formulated, which showed that kriging models based on global approximations are more accurate than Analyzer. However, this improved method is too complexity and time-consuming and can't applied in the engineering (Hui *et al*, 2016; Huang *et al*, 2011; Liu *et al*, 2013).

Differential Evolution algorithm (DE) is a novel swarm intelligence method to search for the optimal result based on the cooperation and competition between different individualities. The DE algorithm is regarded as one of the best optimization method. Some experiments over several numerical benchmarks show that DE performs better than the Genetic algorithm (GA) or the Particle Swarm Optimization (Steentjes *et al*, 2016; Civicioglu *et al*, 2013). In order to improve the DE's performance and solve the problems such as convergence speed and time-consuming, some improved DE algorithm is proposed (Sharma *et al*, 2014; Padhye *et al*, 2015). Considering the robustness and briefness, the DE method gradually gets more and more concentration among pattern recognition, nonlinear optimize control, mechanical engineering and so on (Neri *et al*, 2010; Weber *et al*, 2010).

Considering current situation, how to establish a more accuracy model effectively and quickly is the main problem. Due to the disadvantage of existed approaches, an improved DE based method to estimate

kriging model's parameters is presented in this paper. DE algorithm is an outstanding method to search for the global optimization parameters with high speed. The approximate optimal parameters can be obtained by this algorithm and be used to build an optimal model. At the end of paper, an engineering example of electromagnetic railgun exterior ballistic data is examined to demonstrate the superiority of kriging model based on DE algorithm.

2 Theory of kriging model

Considering a simulation system, a set of m design sites and the output response sites be expressed as

$$\begin{aligned} S &= [s_1\ s_2 ... s_m]^T, s_i \in \mathbb{R}^n \\ Y &= [y_1\ y_2 ... y_m]^T, y_i \in \mathbb{R}^q \end{aligned} \tag{1}$$

Where $s_i = \{s_{i1}, s_{i2,} ... s_{in}\}$ and $y_i = \{y_{i1}, y_{i2,} ... y_{iq}\}$ is the *i*th order in the experiments. The kriging model regards the deterministic response $y(x)$ for an *n* dimensional input $x \in \mathbb{R}^n$ as a combination of a regression model and a random function

$$y(x) = f(x)^T \beta + z(x) \tag{2}$$

Where, $f(x)$ is a vector component of 1, *x* and the other high order items, β is the regression coefficient, $z(x)$ is the random function, which is assumed to have a zero mean and σ^2 variance. The covariance between *n* dimensional inputs x_i and x_j be expressed as

$$Cov\left(z(x_i), z(x_j)\right) = \sigma^2 R(\theta, x_i, x_j) \tag{3}$$

Where, $R(\theta, x_i, x_j)$ is the correlation function with specified parameters θ.

For the set *S* of design sites, a $m \times p$ design matrix *F* is constructed with $F_{ij} = f_j(s_i)$

$$F = \left[f(x_1), f(x_2), ..., f(x_m) \right]^T \tag{4}$$

Furthermore, a correlation matrix *R* between the design sites be defined as $R_{ij} = R(\theta, s_i, s_j)$

$$r(x) = \left[R(\theta, s_1, x), R(\theta, s_2, x), ..., R(\theta, s_m, x) \right]^T \tag{5}$$

The estimation value of output response $\hat{y}(x)$ is a linear combination of response in the design sites

$$\hat{y}(x) = \lambda(x)^T Y \tag{6}$$

The kriging model regards $\hat{y}(x)$ as an optimal linear unbiased estimation of output response $y(x)$. Then, the problem of kriging interpolation can be transformed to an optimization problem

$$\begin{aligned} &\min\ E\left\{ \hat{y}(x) - \lambda(x)^T Y \right\}^2 \\ &s.t.\ \ E\left(\hat{y}(x) - \lambda(x)^T Y \right) = 0 \end{aligned} \tag{7}$$

The solution to the optimization problem above is

$$\hat{y}(x) = f^T(x)\hat{\beta} + r^T(x) R^{-1}\left(Y - F\hat{\beta}\right) \tag{8}$$

Where, the least squares solution of parameter $\hat{\beta}$ is

$$\hat{\beta} = \left(F^T R^{-1} F\right)^{-1} F^T R^{-1} Y \tag{9}$$

The Mean Square Error (MSE) of $\hat{y}(x)$ is

$$\varphi(x) = \sigma^2 \left\{ 1 - \left[f^T(x)\ \ r^T(x) \right] \begin{bmatrix} 0 & F^T \\ F & R \end{bmatrix} \begin{bmatrix} f(x) \\ r(x) \end{bmatrix} \right\} \tag{10}$$

3 Kriging model based on DE algorithm

In this section, the improved method based on DE algorithm is described in detail. Firstly, the basic theory and operations are introduced to support the application in the kriging model. Then, the second part is to assess the kriging model's performance, and lead to an optimization problem. Finally, DE method is applied to solve the optimization problem, and the flowchart of DE based kriging model is used to show the process.

3.1 Theory of DE algorithm

Differential Evolution (DE) is a well-known and simple approach for global optimization, which consists of three basic operations: mutation, crossover and selection. The compute process of DE algorithm is familiar with GA method and can be summarized in the following steps:

- Initialization: set the algorithm's parameters including population members *NP*, variable number *D*, mutagenic factor *F* and crossover probability *CR*, then the initial population

$\left\{ x_{i,j}(0) \mid x_{i,j}^l \le x_{i,j}(0) \le x_{i,j}^u, i = 1, 2, ..., NP, j = 1, 2, ..., \mathrm{D} \right\}$

can be generated by

$$x_{i,j}(0) = x_{i,j}^l + rand(0,1) \times \left(x_{i,j}^u - x_{i,j}^l \right) \tag{11}$$

Where $x_{i,j}(0)$ is the *j*th variable among the *i*th individuality in the 0th generation, $x_{i,j}^u$ and $x_{i,j}^l$ are the upper and lower bound of $x_{i,j}$, respectively. The variable rand(0,1) is the random number of uniform distribution in the interval $(0,1)$.

- Mutation: mutation is a basic operation in the DE algorithm. Which is the largest difference compared with the GA method. The mutation operation can be described as

$$v_i(t+1) = x_{r_1}(t) + F \times \left(x_{r_2}(t) - x_{r_3}(t) \right) \tag{12}$$

Where, r_1, r_2, r_3, i are not equal with each other and $r_1, r_2, r_3 \in \{1, 2, ..., NP\}$, $v_i(t+1)$ is the (*t*+1)th generation mutation individuality, $x_i(t)$ is the *i*th individuality in the *t*th generation.

- Crossover: for the given individuality $x_i(t)$, it's necessary to use the operation of crossover to generate new experiment individuality $u_i(t)$. The equation is

$$u_{i,j}(t+1)=\begin{cases} v_{i,j}(t+1), \text{rand}(j)\le CR \text{ or } j=randn(i) \\ x_{i,j}(t), \text{rand}(j)>CR \text{ or } j\ne randn(i) \end{cases} \quad (13)$$

Where $\text{rand}(j)\in[0,1]$ is the random number of uniform distribution, j is the jth variable of the individuality, $randn(i)\in\{1,2,...,D\}$.

- Selection: the greedy strategy is applied in the DE algorithm to search for the best individuality in a population between $u_i(t)$ and $x_i(t)$, the operation can be expressed as

$$x_i(t+1)=\begin{cases} u_i(t+1), f(u_i(t+1))< f(x_i(t)) \\ x_i(t) \quad , f(u_i(t+1))\ge f(x_i(t)) \end{cases} \quad (14)$$

After iterating the operations above for many times, it is easy to get the optimal vector in the solving space, which will be used in the optimization of kriging model's parameters in the next step.

3.2 Kriging interpolation based on DE algorithm

In general, the quality of a model can be measured by two aspects: 1) the accuracy of data in the design sites. 2) Accuracy in predicting the output response at the estimating points. However, due to the unbiased predictor of kriging model, which means the estimation in the design sites is exactly, the first measurements don't need to consider in this work. And the second aspect to calculate the accuracy at the estimating points is very significant for kriging model.

For a simulation system, the input set of predictor is assumed as $P=[p_1\ p_2...\ p_l]^T, p_i\in\mathbb{R}^n$ and the system real output response and kriging predictor can be expressed as

$$\begin{aligned} Y_p &= [y_{p_1}\ y_{p_2}...\ y_{p_l}]^T, y_i\in\mathbb{R}^q \\ \hat{Y}_p &= [\hat{y}_{p_1}\ \hat{y}_{p_2}...\ \hat{y}_{p_l}]^T, \hat{y}_i\in\mathbb{R}^q \end{aligned} \quad (15)$$

The sum of squared prediction errors is

$$\begin{aligned} SS_T &= \sum_{i=1}^{n}\left(y_{p_i}-\bar{y}_p\right)^2 \\ &= \sum_{i=1}^{q}\left(\hat{y}_{p_i}-\bar{y}_p\right)^2+\sum_{i=1}^{n}\left(\hat{y}_{p_i}-y_{p_i}\right)^2 \\ &= SS_R+SS_E \end{aligned} \quad (16)$$

The accuracy of kriging model is defined as

$$R^2_{fit}=1-SS_E/SS_T, R^2_{fit}>0 \quad (17)$$

For the definition of kriging model accuracy, $R^2_{fit}\in(0,1)$, the value of R^2_{fit} determine the fitting accuracy. In order to get a higher accuracy, the value of R^2_{fit} is as large as it can be. Furthermore, the sum of squared errors of prediction is a fixed value for a specific set of predictors, which means the value of SS_E is as small as it can be. Thus, the model assessment can be transformed to an optimization problem

$$\begin{aligned} &\min\ \sum_{i=1}^{n}\left(\hat{y}_{p_i}-y_{p_i}\right)^2 \\ &s.t.\ \ \hat{y}_{p_i}=f^T(p_i)\hat{\beta}+r^T(p_i)R^{-1}\left(Y-F\hat{\beta}\right) \end{aligned} \quad (18)$$

For a given form of correlation function, the key factor that has a highest influence on the fitting accuracy is the parameters' estimation. From the objective function above, it's obvious that the kriging model accuracy is related with the parameters β and R. There are two methods to estimate the parameters, which are maximum likelihood estimation (MLE) and cross validation (CV). The MLE method was used in this paper and estimation results of β and σ^2 is

$$\begin{aligned} \hat{\beta} &= \left(F^TR^{-1}F\right)^{-1}F^TR^{-1}Y \\ \hat{\sigma}^2 &= \frac{1}{n}(y-F\beta)^T R^{-1}(y-F\beta) \end{aligned} \quad (19)$$

It's clearly that the parameters β and R are related with the estimate value of $\theta=\{\theta_1,\theta_2,...,\theta_n\}$, which transform the objective function's optimization to estimate the parameters θ. Two computational problems often exist when estimating the parameters using traditional method: *1)* the maximum likelihood estimate of the parameters may be multimodality and hardly to find the optimal result. Therefore, the selection of initial value of parameters has a strong influence on the estimation result. *2)* The method above is very suitable for approximating low dimensional model, and has a poor effect of high dimensional one. Considering the robustness and practicability of DE algorithm to solve nonlinear, non-differentiable, multi-extremum and high dimensional problems, an improved kriging model based on DE algorithm is proposed in this paper.

3.3 Process of kriging Model based on DE algorithm

The kriging model's parameters $\theta=\{\theta_1,\theta_2,...,\theta_n\}$ are the variable need to optimize which is determined by the system input. Due to the optimization's purpose is to improve the model fitting accuracy, (11) is utilized to act as the objective function of DE algorithm. The process to optimize kriging model's parameters is presented as show in Figure 1, which can summarize in three aspects:

- Setting DE algorithm parameters: set the objective function with (18) to evaluate the efficiency of DE algorithm. Depend on the research target and compute scope, initialization the DE algorithm's parameters and

carry out the operation of crossover, mutation and selection.

- Data preprocessing: generally, the establishment of kriging model is based on the assumption that the system's input and output data satisfy the normal distribution $N(0,1)$. So it's significant to verify and normalize data's normalization before constructing the kriging model. As stated in Sec. II, the system input samples and output response can be expressed as

$$\begin{aligned} S &= [s_1\ s_2 ... s_m]^T, s_i \in \mathbb{R}^n \\ Y &= [y_1\ y_2 ... y_m]^T, y_i \in \mathbb{R}^q \end{aligned} \tag{20}$$

For simple, data normalization can be described as follows:

$$\begin{aligned} I_i &= \left(s_{:,i} - \overline{s_{:,i}}\right) / \sigma_{s_{:,i}}, i = 1,2,.....,n, \\ O_j &= \left(y_{:,j} - \overline{y_{:,j}}\right) / \sigma_{y_{:,j}}, j = 1,2,...,q \end{aligned} \tag{21}$$

Where, $\overline{s_{:,i}}$ and $\overline{y_{:,j}}$ is the mean of ith input and jth output variable, respectively; $\sigma_{s_{:,i}}$ and $\sigma_{y_{:,j}}$ is the mean variance of ith input and jth output variable, respectively. The calculate equation is

$$\begin{aligned} \overline{s_{:,i}} &= 1/m \sum_{i=1}^{m} s_{ij}, \ \overline{y_{:,j}} = 1/m \sum_{i=1}^{m} y_{:,j} \\ \sigma_{s_{:,i}} &= \sqrt{\sum_{i=1}^{m} (s_{ij} - \overline{s_{:,i}})^2}, \ \sigma_{y_{:,j}} = \sqrt{\sum_{i=1}^{m} (y_{ij} - \overline{y_{:,j}})^2} \end{aligned} \tag{22}$$

After normalization, the mean and variance of the input and output data are 0 and 1, respectively.

- Model computation and optimization: select the specific kriging model including type of correlation function and form of regression function, and then optimize the kriging model's parameters based on DE algorithm, which is stated in Sec. II. As shown in Figure 1, the improved method can be divided into two aspects: the DE algorithm global search and kriging model establishment.

From the process above, it's obvious that the DE algorithm is very suitable for the kriging model's optimization. With the DE method, the best parameters can be gained to establish an optimal kriging model.

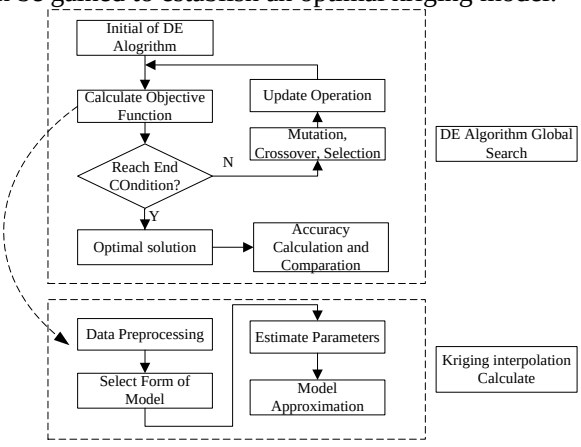


Figure 1. Flow chart of the kriging based on DE method.

4 An engineering example

To demonstrate the validity of kriging model based on DE algorithm, EM railgun exterior ballistic simulation data is taken for example. Due to the complexity of railgun ballistic, the flight range of projectile suffers from a variety of factors, which have a different degree of effect, and coupled with each other. Thus, during the research of projectile's performance, it's crucial to study the relation between projectile's flight range and a set of factors. The linear regression model is the ordinary method to solve this problem. However, the different factors may interaction with each other, and hardly to draw an expression with the projectile's range. Furthermore, the classic regression is the maximum likelihood estimation of sampling points and hardly to solve the problem of multivariable and multimodality. Thus, the kriging model based on DE algorithm is used to establish the model between range and factors.

In this paper, six separate factors including projectile mass, launch velocity, launch angle, deflection angle, wind velocity in x and z direction, are considered in the paper to establish the relationship with flight range using kriging model. Meanwhile, DE algorithm was utilized to optimize the kriging model's parameters and get a higher accuracy.

4.1 Setting DE algorithm parameters

In order to improve the efficiency of DE algorithm and speed up the algorithm convergence, the parameters of DE algorithm set as follows:

The size of population is 30. Mutagenic factor is 0.5, and crossover probability factor is 0.7. Considering the optimal parameters set $\theta = \{\theta_1, \theta_2, ..., \theta_n\}$ represent weight of each dimension, it's important to restrict the span of $\theta_i, i = 1,2...,n$. In general, the range of θ_i is set as $[0.1d_{\min}, 10.0d_{\max}]$, where $d_{\min}$ and $d_{\min}$ represents the minimum and maximum distance of ith input parameter sets, respectively.

4.2 Data preprocessing

Before establishing the kriging model, it's obvious to preprocess the flight range data. The normalization verification result is demonstrated in Figure 2 and 3, which including two graphs: the left and right one is the bar graph of flight range and result of normality test, respectively. As the verify result show, the disperse position of range data coincide with the standard normalization reference line. Furthermore, it's necessary to take the range data with normalization operation, which is obtained from (12) and (13). Thus, the railgun data meets the requirement of kriging model construction after preprocessing.

4.3 Model computation and optimization

After accomplishing the data preprocessing, the last step is to compare the model accuracy by two aspects: 1) the approximation accuracy of kriging model based on DE algorithm and ordinary one, 2) the accuracy of different forms of correlation function and regression function $f(x)$. In order to demonstrate the superiority of improved method, three types of correlation function and five forms of regression function $f(x)$ is adapted in the DE method in this paper. The types of correlation function consist of Exp, Gauss and Linear and the forms of regression function are summarized in Table 1.

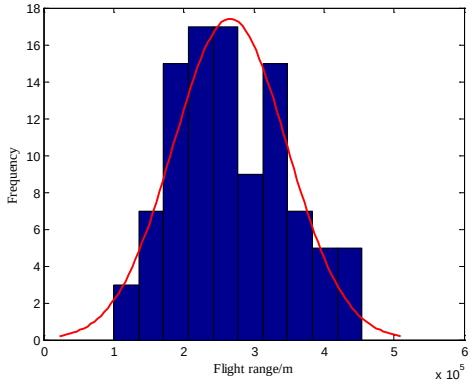


Figure 2. Histogram of the range data.

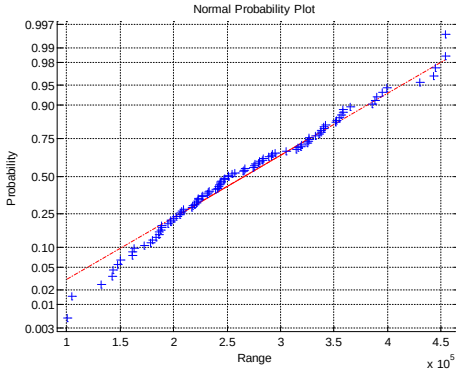


Figure 3. Normalized test result.

Given the different types of correlation function and forms of regression function in the kriging model, the DE algorithm is used to search for the optimization of kriging model's parameters. The sum error data of DE based kriging model and traditional one are listed in Table 2 and 3. In order to have an intuitive grasp of these two methods, a histogram of kriging model error is showed in the Figure 4, where the left and right bar is the improved method and traditional one, respectively. In the Figure 4, the horizontal and longitudinal ordinates of histogram represent the forms of $f(x)$ and sum of approximation error. The three colors indicate three type of correlation function Exp, Gauss and Linear. At the aspect of DE algorithm convergence speed, the kriging model which consists of the fourth form of regression function and Gauss correlation function is examined to reveal the iteration process in Figure 5.

Table 1. Forms of the regression function $f(x)$.

Indication	*Forms of* $f(x)$
1	1
2	$[1, x_1, x_2, ..., x_n]$
3	$[1, x_1, x_2, ..., x_n, x_1^2, x_1x_2, x_1x_3..., x_n^2]$
4	$[1, x_1, x_2, ..., x_n, x_1^2, x_2^2, ..., x_n^2]$
5	$[1, x_1, x_2, ..., x_n, x_1x_2, x_1x_3..., x_{n-1}x_n]$

Table 2. Sum errors of kriging model based On DE.

Indication	*EXP*	*GAUSS*	*LIN*
1	5.58e4	4.99e4	4.46e4
2	3.065e4	1.99e4	2.52e4
3	1.86e4	5.75e3	1.59e4
4	1.38e4	6.52e3	1.06e4
5	3.51e4	2.23e4	3.2e4

Table 3. Sum errors of traditonal kriging model.

Indication	*EXP*	*GAUSS*	*LIN*
1	9.77e4	1.08e5	2.34e5
2	3.11e4	2.53e4	3.43e4
3	1.93e4	1.2e4	3.29e4
4	1.47e4	8.49e3	2.72e4
5	3.77e4	3.21e4	4.05e4

As show in Figure 4, the kriging model based on DE algorithm is prior to the traditional one in the model's accuracy. For the different types of correlation function, the Gauss model has a significant advantage than the other models. Meanwhile, the fourth type of regression function has the highest accuracy in the five forms of regression function. So it's essential to compare different forms of regression function to select the highest accuracy one.

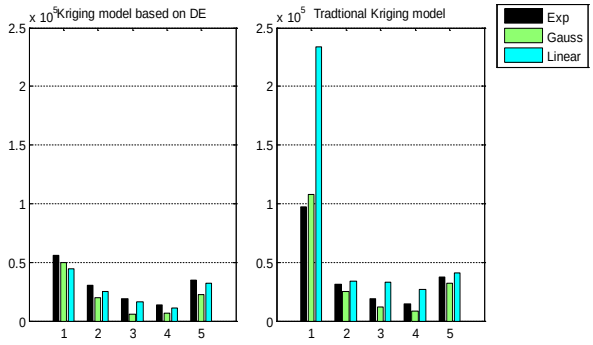


Figure 4. Histogram of the kriging model error.

Owing to the accuracy of kriging model has a slight relation with complexity of regression function. Figure 5 also prove the high speed of DE algorithm

convergence, it's obvious that the approach could find the optimization result quickly and effectively. Thus, the kriging model based on DE algorithm improves the approximate on accuracy and has a valuable application in engineering.

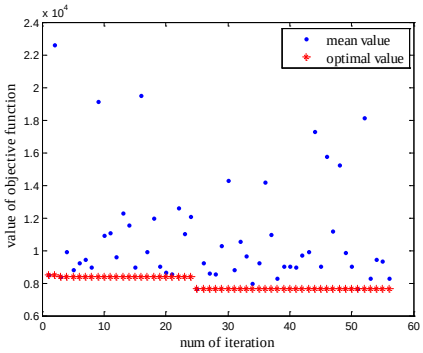


Figure 5. Iteration process of the DE optimization.

5 Conclusions

The parameters estimation and approximate form selection to obtain a higher accuracy model is a basic existed problem in the research of kriging interpolation. In the traditional method, the model parameters selection always depends on users. In order to solve the problem and get an optimal model, an improved kriging model based on DE algorithm to approximate simulation model was presented in this article. The research of kriging model proves that correlation function's parameters have a strong influence on the fitting accuracy. Then the DE algorithm establishes the objective function by fitting accuracy to optimize the kriging model's parameters. The EM railgun exterior ballistic data was taken for instance to demonstrate the priority of improved method. Three forms of correlation function and five types of regression model are used to compare the approximation accuracy of kriging model based on DE algorithm with traditional one. The simulation results show that the kriging model based on DE algorithm has the higher accuracy, and a fine prospect of engineering application.

There is a need for more research on the improved method. Although the DE algorithm perform well in kriging model, it's essential to have deep research of the DE parameters such as population members NP, variable number D and mutagenic factor F, which determine the search efficiency. So the future work can be concentrated on parameters selection strategy to obtain a higher speed.

Acknowledgements

This work is supported by National Science Foundation of China (No. 61374164).

References

P. Civicioglu. Backtracking Search Optimization Algorithm for Numerical Optimization Problems. *Applied Mathematics and Computation*, 219(15):8121-8144, 2013. doi:10.1016/j.amc.2013.02.017

Z. Hui, C. Wang, and J. Chen. Optimal Design of Aeroengine Turbine Disc based on Kriging Surrogate Models. *Computers & Structures*, 89(1):27-37, 2011. doi: 27-37.10.1016/j.compstruc.2010.07.010

Z. Hui, Y. Hu, and Y. Yevenyo. An Improved Morphological Algorithm for Filtering Airborne LiDAR Point Cloud Based on Multi-Level Kriging Interpolation. *Remote Sensing*, 8(5):1-16, 2016. doi: 10.3390/rs8010035

J. P. C. Kleijnen. Kriging Metamodeling in Simulation: A Review. *European Journal of Operational Research*, 192(3):707-716, 2007. doi: 10.1016/j.ejor.2007.10.013

X. L. Chen. Multi Points Updated and Distance Filtered Kriging Surrogate Model. *Application in EOSS Optimization*, 22(1):209-213, 2013.

S. N. Lophaven, H. B. Nielsen, and J. Søndergaard. Aspects of the matlab toolbox DACE. *Informatics and Mathematical Modelling*. Technical University of Denmark, DTU, 2002.

S. N. Lophaven, H. B. Nielsen, and J. Søndergaard. *DACE-A Matlab Kriging toolbox*, version 2.0. 2002.

J. D. Martin and T. W. Simpson. A Monte Carlo Simulation of the Kriging Model. In *10th AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, 2004-4483, 2004. doi: 10.2514/6.2004-4483

J. D. Martin and T. W. Simpson. Use of Kriging Models to Approximate Deterministic Computer Models. *AIAA Journal*, 43(4):853-863, 2005. doi: 10.2514/1.8650

M. Weber, V. Tirronen and F. Neri, Scale Factor Inheritance Mechanism in Distributed Differential Evolution, *Soft Computing*, 14(11):1187-1207, 2010. doi:1187-120.10.1007/s00500-009-0510-5

F. Neri and V. Tirronen. Recent Advances in Differential Evolution: A Survey and Experimental Analysis. *Artificial Intelligence Review*, 33(1-2):61-106, 2010. doi: 10.1007/s10462-009-9137-2

N. Padhye, P. Mittal, and K. Deb. Feasibility Preserving Constraint-Handling Strategies for Real Parameter Evolutionary Optimization. *Computational Optimization and Applications*, 62(3):851-890, 2015. doi: 10.1007/s10589-015-9752-6

S. Steentjes, M. Petrun, and D. Dolinar. Effect of Parameter Identification Procedure of the Static Hysteresis Model on Dynamic Hysteresis Loop Shapes. *IEEE Transactions on Magnetics*, 52(5):1-4, 2015. doi: 10.1109 /TMAG.2015.2511800

H. Sharma, J. C. Bansal, and K. V. Arya. Self Balanced Differential Evolution. *Journal of Computational Science*, 5(2):312-323, 2012. doi: 10.1016/j.jocs.2012.12.002

S. Volpi, M. Diez, and N. Jgaul. Development and Validation of A Dynamic Metamodel based on Stochastic Radial Basis Functions and Uncertainty Quantification. *Structural and Multidisciplinary Optimization*, 51(2):347-368，2015. doi: 10.1007/s00158-014-1128-5

Simulation of Control Structures for Slug Flow in Riser during Oil Production

Ole Magnus Brastein Roshan Sharma

Department of Electrical Engineering, IT, and Cybernetics, University College of Southeast Norway, Porsgrunn, Norway, {ole.m.brastein, roshan.sharma}@usn.no

Abstract

The occurrence of slug flow is a common problem arising in the oil well riser pipeline. To eliminate such slug flow, various control structures along with state estimation are designed and compared in this paper. Nonlinear model based predictive scheme are compared with classical PI controllers for three different control structures. One of the control structure is based on controlling the mass of the liquid in the riser pipeline, for which, an Unscented Kalman Filter is designed to estimate the mass. The simulation results show that the model based controllers perform relatively better than the classical controllers. Although computationally expensive, the control algorithm used in this paper for model based control still makes it real time implementable.

Keywords: slug flow, oil riser, model based control, PI control, unscented Kalman filter

1 Introduction

In oil well riser pipelines with low-point angle, the liquid column accumulated in the riser above the low point acts as a virtual valve (see Figure 1), alternately blocking and letting the gas produced from the reservoir to flow through the riser. This is due to the hydrostatic pressure exerted at the low point by the liquid column in the riser. The gas produced from the reservoir at first starts to accumulate below the low point. The pressure builds up and reaches to a critical point where the built-up pressure exceeds the hydrostatic pressure drop. This results in a rapid discharge of the accumulated gas to the riser. This large gas bubble/volume pushes up some of the liquid in the riser and out from the choke valve. However, with time, the liquid again starts to accumulate in the riser. The gas pressure at the horizontal flowline starts to build up again and the cycle repeats. This behavior of the fluid flow in the riser is known as slug flow. It is an unstable multiphase flow where oscillation in the production of oil from the reservoir occurs.

Formation of slug flow in oil well riser pipelines has been studied by many other researchers. Various control solutions to avoid the slug flow in oil wells along with the development of mathematical model and controllability analysis can be found at (Storkaas, 2005). Control of riser base pressure for stabilizing the slug flow have been studied by (Aamo et al., 2005; Dalsmo et al., 2005). A simplified model based on first principles modeling for reproducing slugging oscillations of a real oil well was developed by (Meglio et al., 2009). This model was later used by (Meglio et al., 2012b) for designing control structures. Control strategies for slug control and tuning rules was studied by (Godhavn et al., 2005). A nonlinear controller using integrator backstepping approach was used by (Kaasa et al., 2007) to stabilize unstable wells. A review of recent advances in the suppression of the slugging phenomenon by model-based control can be found at (Meglio et al., 2012a). This article gives a clear presentation of the evaluation and comparison of the existing solutions and proposes directions for improvement. A similar type of slugging phenomenon is also observed for gas-lifted oil wells. Stabilization of gas-lift wells by feedback control was studied by (Eikrem, 2006; Imsland, 2002). An insight and understanding into how feedback control can be used to avoid severe slugging, thereby bridging the gap between control and petroleum engineering can be found at (Havre and Dalsmo, 2002).

In this paper, three different control structures/strategies are developed to stabilize the flow in the riser so that the flow does not oscillate and becomes stable. In Section 2, a brief description of the model that captures the slug flow phenomenon in the riser is described. The model is simulated to illustrate its capability of capturing the formation of slug flow in the riser in Section 3. Implementation of an Unscented Kalman Filter(UKF) for estimating the states of the system is provided in Section 4. In Section 5, the formulation of the three different control strategies is presented. The simulation results obtained from the model based controller and the PI controller are clarified, compared and discussed in Section 6. A possibility for the real time implementation and the computational time required by the model predictive controller (MPC) is discussed in Section 7. A brief discussion on the maximum valve opening that can be achieved before the flow becomes unstable again is presented in Section 8. Finally, conclusions are provided in Section 9.

2 Model for slug flow

A widely used mathematical model for representing the slug flow in the oil well risers was developed by (Meglio et al., 2012b) and this model has been used in the present

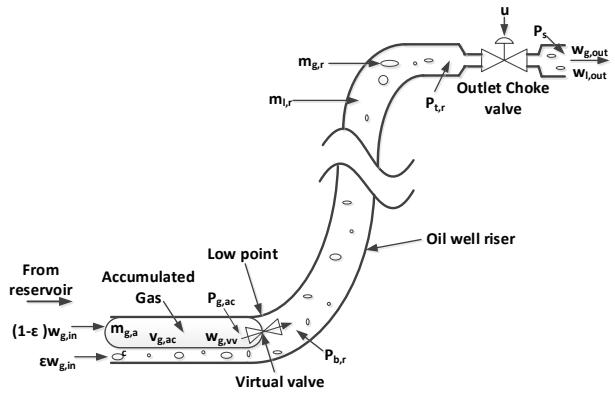


Figure 1. Schematic of fluid transportation in the flow line and riser.

work. Only a brief description of the model is presented in this section and the details about the development of the model can be found at (Meglio et al., 2012b). Let us consider the low point of the riser at the place where the virtual valve is located as shown in Figure 1.

The flowline and the riser are divided into three separate volumes/parts: (i) Volume at the horizontal part of the flowline where the incoming gas from the reservoir accumulates, (ii) Volume of the vertical riser filled with liquid only, and (iii) Volume of the vertical riser filled with gas only. The model is based on the conservation principle where the mass balance is applied to each of the three volumes. The state variables are the mass of the gas accumulated in the horizontal flowline $(m_{g,ac})$, the mass of the gas in the riser $(m_{g,r})$ and the mass of the liquid in the riser $(m_{g,l})$. From the mass balances we obtain,

$$\frac{dm_{g,ac}}{dt} = (1-\lambda)w_{g,in} - w_{g,vv}, \tag{1}$$

$$\frac{dm_{g,r}}{dt} = \lambda w_{g,in} + w_{g,vv} - w_{g,out}, \tag{2}$$

$$\frac{dm_{g,l}}{dt} = w_{l,in} - w_{l,out}. \tag{3}$$

Here, λ denotes the fraction of gas coming from the reservoir that directly flows to the riser, $w_{g,in}$ is the flow rate of the gas entering the riser, $w_{g,vv}$ is the flow rate of the gas through the virtual valve, $w_{g,out}$ is the flow rate of the gas flowing out of the riser through the outlet choke valve, $w_{l,in}$ is the flow rate of the liquid entering the riser and $w_{l,out}$ is the flow rate of the liquid flowing out of the riser through the outlet valve. The algebraic equations involved in the models are listed below. These are taken from (Meglio et al., 2012b) and the details about their development is not provided in this paper.

$$w_{g,vv} = C_{g,vv} max\left(0, (P_{g,ac} - P_{b,r})\right) \tag{4}$$

$$P_{g,ac} = \frac{m_{g,ac}RT}{MV_{g,ac}} \tag{5}$$

$$V_{g,r} = V_r - \frac{\left(m_{l,r} + m_{l,min}\right)}{\rho_l} \tag{6}$$

$$P_{t,r} = \frac{m_{g,r}RT}{MV_{g,r}} \tag{7}$$

$$P_{b,r} = P_{t,r} + \left(m_{l,r} + m_{l,min}\right)\frac{gsin\theta}{A} \tag{8}$$

$$w_{out} = C_{out}u\sqrt{\rho_m\left(P_{t,r} - P_s\right)} \tag{9}$$

$$w_{l,out} = \frac{m_{l,r}}{m_{l,r} + m_{g,r}} w_{out} \approx w_{out} \tag{10}$$

$$w_{g,out} = \frac{m_{g,r}}{m_{l,r} + m_{g,r}} w_{out} \approx \frac{m_{g,r}}{m_{l,r}} w_{out} \tag{11}$$

Here, $C_{g,vv}$ is the valve constant for the virtual valve, $P_{g,ac}$ is pressure of the gas accumulated in the horizontal flowline (upstream the virtual valve), $P_{b,r}$ is the pressure at the bottom of the riser, $P_{t,r}$ is the pressure at top of the riser, P_s is the separator pressure, R is the ideal gas constant, T is the temperature of the fluid in the riser, M is the molar mass of the gas, $V_{g,ac}$ is the volume of the horizontal flowline where the gas accumulates, V_r is the physical volume of the riser, $V_{g,r}$ is the volume of gas in the riser, $m_{l,min}$ is the minimum amount of liquid present in the riser at all times, ρ_l is the density of the liquid, θ is the mean inclination of the riser pipe, A is the cross section of the riser, C_{out} is the valve constant for the outlet choke valve, $u \in [0,1]$ is the valve opening and w_{out} is the total mass flow rate flowing out of the riser through the outlet valve.

3 Simulation for slug flow

The model presented in Section 2 is simulated in MATLAB to observe the occurrence of the slug flow in the riser pipeline. The mass flow rate of the gas and the liquid flowing into the well from the reservoir are considered to be constant. With a nominal valve opening of 0.35 or 35%, the fluid flow in the riser pipe i.e. the outflow from the outlet valve exhibits a slug flow as shown in Figure 2.

The flow of the liquid from the outlet valve oscillates with a time period of about 50 minutes. This is due to the virtual valve that alternately blocks and lets the gas to flow through the riser, thus producing a slug flow. The pressure at the bottom and top of the riser oscillates and this oscillating nature of the pressures in the riser actually creates the slug flow. The average production of oil from the field due to an unstable slug flow is lower than the theoretical steady state (or equilibrium) production. Such unstable slug flow should be controlled or stabilized.

In reality, the process operators choke the outlet valve manually to stabilize the slug flow. The slug flow can be stabilized by decreasing the opening of the output choke valve. A stabilized slug flow results in a non-oscillating pressure at the bottom of the riser. In Figure 3, the choke valve opening is decreased from 45% to 15% in steps. As can be seen from Figure 3, the pressure at the bottom of the riser still keeps on oscillating when the valve opening

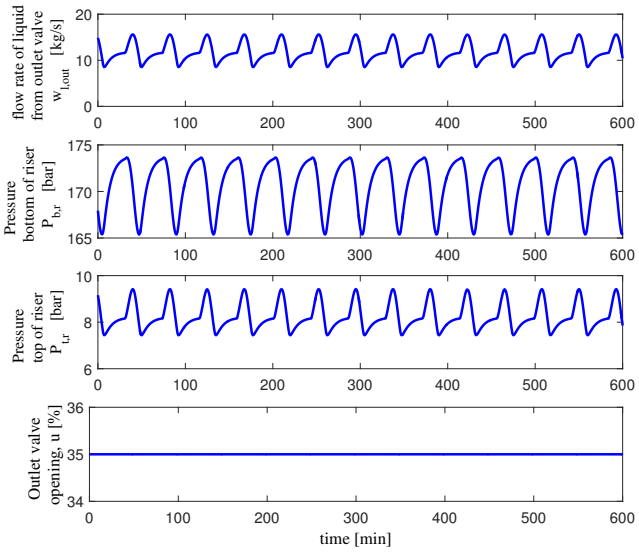


Figure 2. Slug flow in the riser for u = 35%.

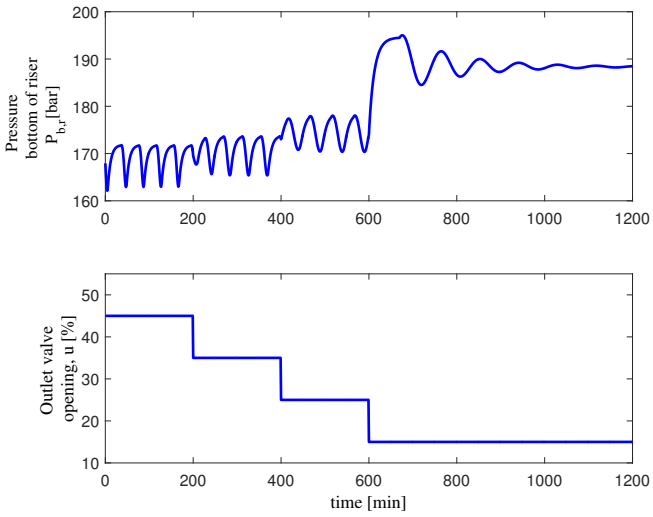


Figure 3. Variation of pressure at the bottom of the riser with valve opening.

is reduced from 45% to 25%. At time 600 min, the outlet valve opening is reduced to 15%. With this valve opening, the pressure at the bottom of the riser is stabilized and this results in a stabilized flow of the liquid through the outlet valve. As the valve is slowly choked, at one point, the flow is stabilized. This value of valve opening for which the flow starts to stabilize is called the bifurcation point. Above this point, the flow is unstable and below this point, the flow is stable. From the openloop simulations, the bifurcation point was found out to be around 20% valve opening.

From Figure 3, we can observe that the flow can be stabilized by remaining below the bifurcation point in the stable region. Usually, the bifurcation point corresponds to a low valve opening. With a lower valve opening, the flow rate of the oil produced from the well is also low which is economically not beneficial. Thus, it is of interest to investigate whether the flow can be stabilized by opening the valve in the unstable zone (i.e. by remaining above the bifurcation point) through the use of different control strategies.

4 State estimation

One of the control structure that is explained in detail in Section 5 utilizes the information about the mass of the liquid (which is a state variable) in the riser pipeline. This and the remaining two states of the process cannot be directly measured and hence should be estimated. For this, an Unscented Kalman Filter (UKF) that directly utilizes the nonlinear model of the process is implemented. Details of the UKF is not the main focus of this paper and interested readers are advised to follow (Simon, 2006). In this work, standard algorithm for UKF available at (Simon, 2006) is implemented in MATLAB.

In addition, it is assumed that the pressures at the bottom and top of the riser pipe are measured and available. Figure 4 shows the estimated states of the system and Figure 5 shows the estimated pressures at the bottom and top of the riser pipeline. The estimated states and the estimated measurements are then used by the control structures for stabilizing the flow.

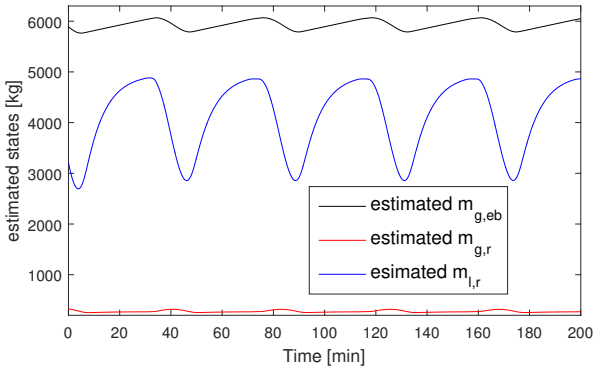


Figure 4. Estimated states by UKF.

5 Control strategies

For regulating the slug flow in the riser pipeline, three different control strategies/structures were developed.

- The first control structure stabilizes the slug flow by controlling the pressure at the bottom of the riser i.e. by controlling $P_{b,r}$ to a set point.
- The second control structure stabilizes the slug flow by controlling the pressure drop in the riser i.e. by controlling $\triangle P = (P_{b,r} - P_{t,r})$ to a set point.
- The third control structure stabilizes the slug flow by controlling the total mass of the liquid in the riser i.e. by controlling $m_{l,r}$ to a setpoint.

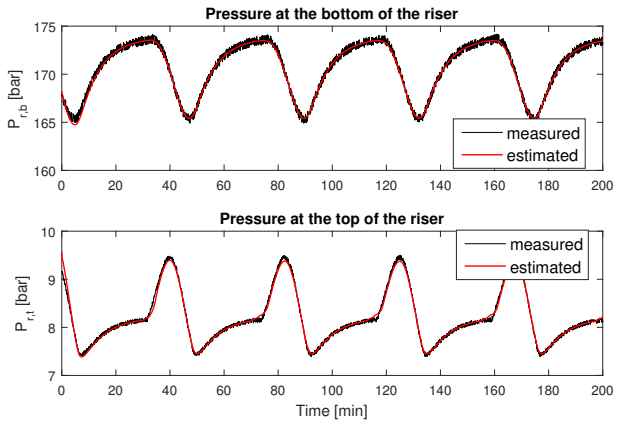


Figure 5. Estimated pressures at the bottom and top of the riser.

Each of these control strategies were implemented as a model based controller using model predictive control (MPC) scheme as well as using a standard Proportional-Integral (PI) control scheme.

5.1 Model predictive control

For designing a nonlinear model predictive controller, consider the following nonlinear objective function,

$$\min_{\triangle u_c} f(\triangle u_c) = \sum_{k=1}^{N_p} \left(XX_k - XX_k^{ref}\right)^T P_k \left(XX_k - XX_k^{ref}\right) + \sum_{k=1}^{N_c} (\triangle u_k)^T R_k (\triangle u_k) \tag{12}$$

Here, XX_k is the variable to be controlled and XX_k^{ref} is its reference value depending on the choice of the control structure. $XX_k = P_{b,r}$ for control structure 1, $XX_k = \triangle P = (P_{b,r} - P_{t,r})$ for control structure 2 and $XX_k = m_{l,r}$ for control structure 3. N_p is the prediction horizon length and N_c is the control horizon length. P_k is the weighting factor for the set point error and R_k is the weighting factor for the control deviation. $\triangle u_k = u_k - u_{k-1}$ is the rate of change of control action.

The choke valve opening should be between 0 and 1, i.e. the constraint in the control input is,

$$0 \leq u_k \leq 1 \tag{13}$$

In practice, the choke valves are opened in smaller steps, and larger abrupt changes in its opening is usually avoided. Consider that the choke valve can be opened or closed by only 0.5% per second i.e.

$$-0.5\% \leq \triangle u_k \leq 0.5\% \tag{14}$$

(13) and (14) together with the model of the process form the constraints for the optimization problem. Moreover, in order to improve the speed of computation without loosing any control dynamics, the prediction horizon was grouped into four groups.

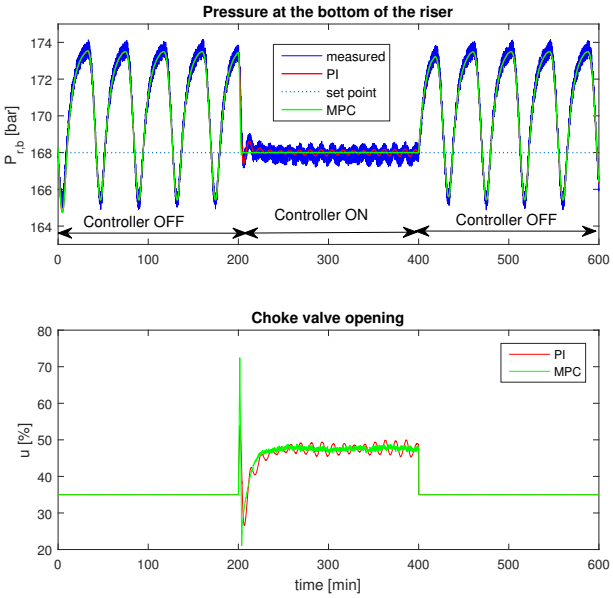


Figure 6. Pressure at the bottom of the riser with control structure 1.

5.2 PI control

A standard expression for the PI controller in the deviation form can be written as,

$$\triangle u_k = K_p \left(1 + \frac{dt}{2T_i}\right) e_k - K_p \left(1 - \frac{dt}{2T_i}\right) e_{k-1}, \tag{15}$$

with

$$e_k = XX_k^{ref} - XX_k, \tag{16}$$

corresponding to the three control structures respectively. The conditions fulfilling (13) and (14) were implemented together with (15) and (16). Here, K_p is the proportional gain of the controller, T_i is the integral time constant and dt is the sampling time taken to be 5 seconds.

6 Simulation results and discussion

6.1 Control structure I

The set point for the pressure at the bottom of the riser was chosen to be 168 bar. Figure 6 and Figure 7 show the simulation results both with the model predictive controller and with the PI controller. Each controller was turned on between the time interval of 200 and 400 minutes. Both controllers were able to stabilize the outflow (see Figure 7) however, the performance of the model predictive controller was better than the PI controller. Small oscillations of the pressure at the bottom of the riser and in the choke valve opening were seen with the PI controller. However with MPC, such small oscillations were completely eliminated.

From the openloop simulations it is known that the bifurcation point of the valve opening is around 20%. However, with this control structure, the valve remains opened

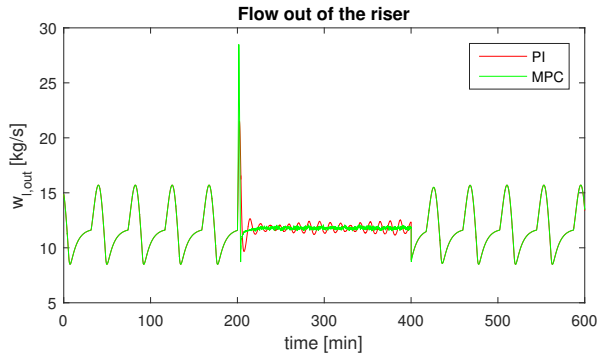


Figure 7. Stabilization of outflow with control structure 1.

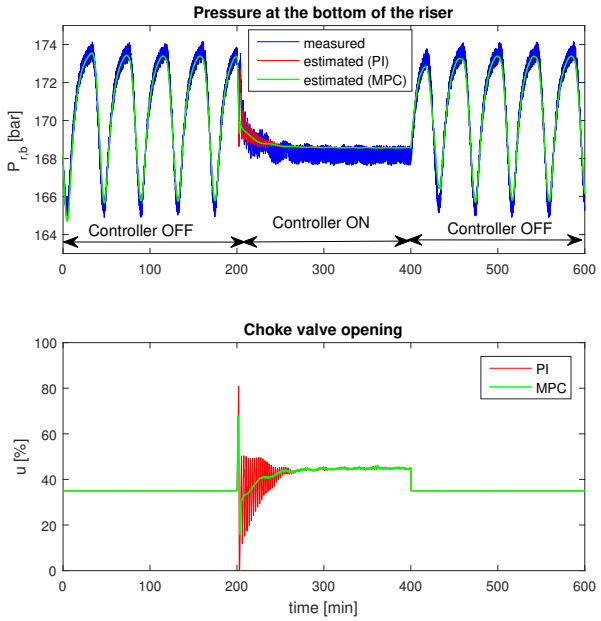


Figure 8. Pressure at the bottom of the riser with control structure 2.

at around 48% which is in the unstable region of the valve opening. This clearly indicates that it is possible to stabilize the flow flowing in the riser pipeline while still staying at the unstable region of the valve opening. This is an added benefit with respect to the operation of the process: the more the valve opening, the more is the amount of oil flowing out of the well (economically more beneficial).

6.2 Control structure II

For control structure II, the set point for the pressure difference between the bottom and the top of the riser was taken to be 161 bar. Figure 8 and Figure 9 show the simulation results both with the model based controller and with the PI controller. Both control schemes were able to stabilize the slug flow. The simulation results show that the bottom hole pressure can be indirectly stabilized by controlling the pressure drop over the riser. However, the model based controller outperforms the PI controller. High frequency oscillations in the valve opening

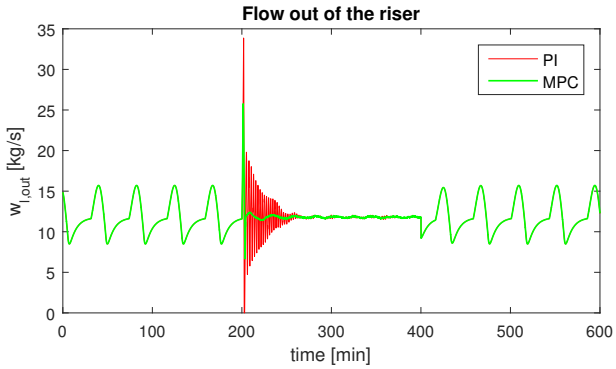


Figure 9. Stabilization of outflow with control structure 2.

(and in the flow through the choke valve) were observed at the beginning of the control action with the PI controller, while such oscillations were completely suppressed by the model predictive controller. The control scheme utilizes the estimated values calculated by the UKF. If the controller uses the measurement of the pressures directly for the calculation of control actions, the dynamics become even more oscillatory (with higher frequency oscillations). Therefore for this control structure, proper tuning of the UKF is very essential. Compared to control structure 1, more weight (10 times more) was put to the measurement noise covariance matrix during the implementation of UKF.

6.3 Control structure III

With this control structure, the mass of the liquid in the riser was controlled to a set point of 3800 kg. Since the mass of the liquid in the riser has a significant and a direct effect on the pressure at the bottom of the riser, controlling it allows the bottom hole pressure to be controlled indirectly. Figure 10 and Figure 11 show the simulation results. It is very clear that for this control structure, the PI controller does not stabilize the flow properly. The model predictive controller outperforms the PI controller. Although the amplitude of the oscillation for the pressure at the bottom of the riser was lowered with the PI controller (which is better than without any control at all), but at the same time the frequency of the oscillation was increased. The valve openings oscillated periodically and the flow of the fluid through it oscillated with a more higher frequency than before (without any control). However, with the model predictive controller, the control action was superior without any oscillations. The flow of fluid out of the riser and the pressure at the bottom of the riser were very stable.

7 Computational time for MPC

Model predictive controller is a computationally heavy algorithm. At each iteration, a constrained nonlinear optimization problem is solved. In this work, a prediction horizon of 30 samples with a sampling time of 5 sec which equals to 150 sec was used. The control inputs

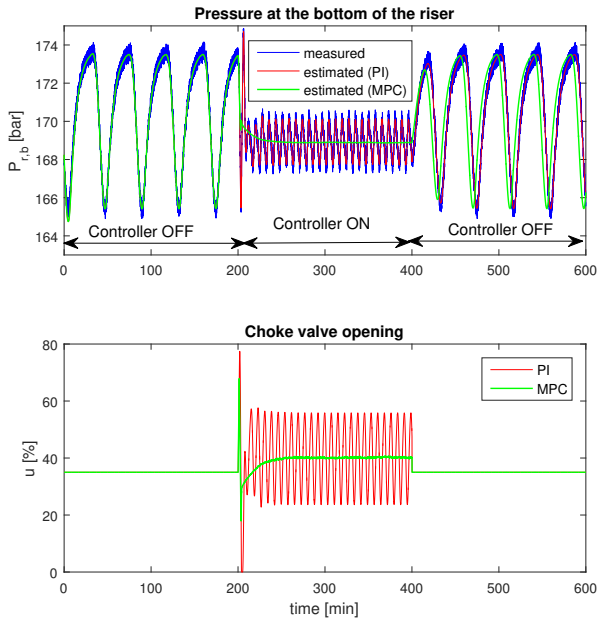


Figure 10. Pressure at the bottom of the riser with control structure 3.

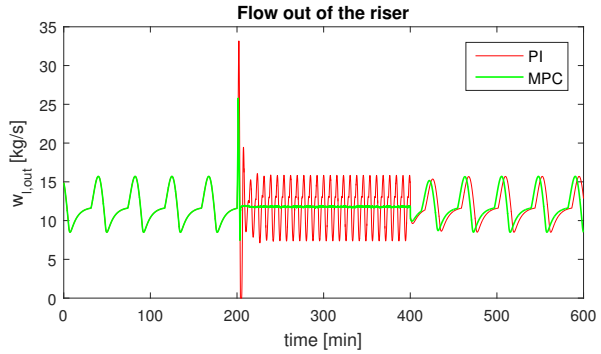


Figure 11. Stabilization of outflow with control structure 3.

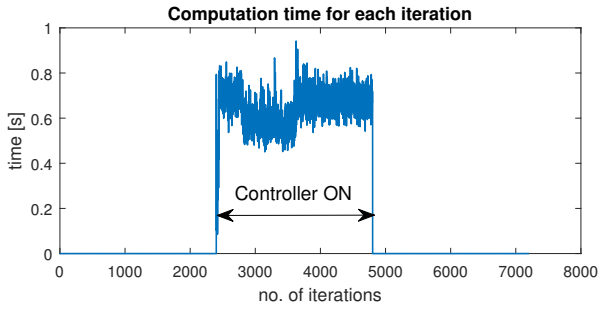


Figure 12. Computational time required by the model predictive controller for structure 2.

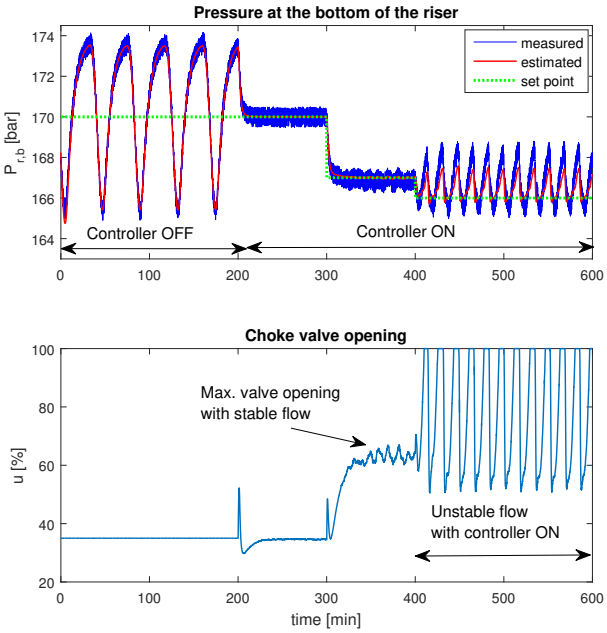


Figure 13. Maximum valve opening for stable flow with control structure 1.

were grouped into four groups. Thus, instead of optimizing 30 unknowns at each iteration, only 4 unknowns are optimized. This significantly reduces the computational time without deteriorating the control action. For all the three control structures, each iteration could be solved in less than a second, and for the chosen sampling interval, this means that the algorithm can be easily implemented for real time application. For an illustration, the computational time required by the MPC algorithm for control structure 2 is shown in Figure 12. A normal computer with 2.50 Ghz processor and 4 GB RAM was used for the simulations.

8 Maximum valve opening

The choice of the setpoint(s) for all the three control structures were arbitrary and can be freely chosen. However, the choice of the setpoint influences the ability of the controller to stabilize the flow. To illustrate this, consider the first control structure. Lowering the value of the setpoint for the pressure at the bottom of the well corresponds to a higher valve opening and hence more flow of oil through the outlet choke valve as shown in Figure 13. With respect to the production of oil, it is economically beneficial to have the outlet valve open as much as possible. In Figure 13, as the setpoint was decreased from 170 bar to 167 bar, the valve opening was increased from 35% to 65% (maximum opening that marks the boundary between the stable and the unstable flow). As the setpoint was further decreased to 166 bar after 400 min, the pressure started to oscillate and the flow became unstable again, even though the controller was still turned on.

The presence of a maximum valve opening before the flow again becomes unstable was also seen with the other two control structures but have not been shown in this paper to save space.

9 Conclusions

From this research work, it was observed that the model predictive control schemes outperform the standard PI controllers for stabilizing the slug flow in oil well riser. The differences are not so significant for the first two control strategies. However, for the third control strategy, the control actions are superior with the model based control. Nonlinear MPC with Unscented Kalman filter can be implemented for real time control of the slug flow in riser. Among the three control structures, with the model predictive control, all the control structures were equally able to stabilize the flow. It is difficult to conclude which of these control structure is the best with the model predictive control. However, with the PI controllers, the first control strategy stabilized the flow better than the remaining two control structures.

References

O. M. Aamo, G. O. Eikrem, H. B. Siahaan, and B. A. Foss. Observer design for multiphase flow in vertical pipes with gas-lift theory and experiments. *Journal of Process Control*, 15:247–257, 2005.

M. Dalsmo, E. Halvorsen, and O. Slupphaug. Active feedback control of unstable wells at the brage field. *Modeling, Identification and Control*, 26(2):81–94, 2005. doi:10.4173/mic.2005.2.2.

G. O. Eikrem. *Stabilization of gas-lift wells by feedback control*. PhD thesis, Norwegian University of Science and Technology, 2006.

J. M. Godhavn, M. P. Fard, , and P. H. Fuchs. New slug control strategies, tuning rules and experimental results. *Journal of Process Control*, 15:547–557, 2005.

K. Havre and M. Dalsmo. Active feedback control as a solution to severe slugging. volume 17, New Orleans, Lousiana, 2002. Society of Petroleum Engineers. doi:10.2118/79252-PA.

L. S. Imsland. *Output Feedback and Stabilization and Control of Positive Systems*. PhD thesis, Norwegian University of Science and Technology, Department of Engineering Cybernetics, Trondheim, Norway, 2002.

Glenn-Ole Kaasa, Vidar Alstad, Jing Zhou, and Ole Morten Aamo. Nonlinear model-based control of unstable wells. *Modeling, Identification and Control*, 28(3):69–79, 2007.

F. D. Meglio, G. O. Kaasa, N. Petit, and V. Alstad. Reproducing slugging oscillations of a real oil well. In 49^{th} *IEEE Conference on Decision and Control*, pages 4473–4479, Hilton Atlanta Hotel, Atlanta, GA, USA, December 15-17 2009.

F. D. Meglio, G. O. Kaasa, N. Petit, and V. Alstad. Model-basedcontrolofslugging:advancesand challenges. In *2012 IFAC Workshop on Automatic Control in Offshore Oil and Gas Production*, pages 109–115, Norwegian University of Science and Technology, Trondheim, Norway, May 31- June 1 2012a.

F. D. Meglio, N. Petit, V. Alstad, and G. O. Kaasa. Stabilization of slugging in oil production facilities with or without upstream pressure sensors. *Journal of Process Control*, 22: 809–822, 2012b. doi:10.1016/j.jprocont.2012.02.014.

D. Simon. *Optimal State Estmation: Kalman, H_∞ and Nonlinear Approaches*. John Wiley & Sons, Inc., 2006.

E. Storkaas. *Control Solutions to Avoid Slug Flow in Pipeline-riser Systems*. PhD thesis, Norwegian University of Science and Technology, 2005.

Security Threats and Recommendation in IoT Healthcare

Cansu Eken, Hanım Eken

Computer Science, Ankara University, Turkey, cansueken21@gmail.com
Türksat,Turkey,eken.hanim@gmail.com

Abstract

The Internet of Things (IoT) devices have become popular in recent year. All devices connect network and communicate each other. Therefore all devices become smart. They are used for some systems such as e-Health, e-Energy, e-Home, smart city, smart car etc. IoT device collect data for systems in order to analyze data and give right decision. Thus, attackers attack IoT systems. This paper gives an introduction to IoT healthcare systems and applications, the related security and privacy challenges. This paper tends to analyze the security threats in different layers of the loT, and give recommendation owing to provide security and privacy.

Keywords: internet of things (IoT), body area network, wearable devices, IoT healthcare systems, security of IoT, privacy, information security

1 Introduction

In recent years, many devices and goods are used in life and these devices and goods communicate with each other. These devices are radio-frequency identification (RFID) tags, mobile phones, mobile devices, sensors, actuators, etc. The term ‘‘Internet-of-Things’’ (IoT) is broadly used to refer to physical objects or "things" connect each other and exchange data on network. The term, internet of things (IoT) was first proposed in 1998 (Weber, 2010).

In the year of 2005, International Telecommunication Union (lTU) released an annual report on "Internet of Things. This report includes the development of information and communication technology and the future of smart devices, communicating with each other (ITU, 2005). Physical objects or "things” become smart by connecting to each other. In later years, physical objects or "things” in every aspect of our lives connect to internet and exchange to data. Therefore, ‘‘Internet-of-Things’’ (IoT) devices are used many sectors.

According to Gartner, Inc., there will be nearly 26 billion devices on the Internet of Things by 2020. Figure 1 presents a list of countries by IoT devices online per 100 inhabitants as published by the OECD in 2015.

Rank	Country	Devices online	Relative size
1	South Korea	37.9	
2	Denmark	32.7	
3	Switzerland	29.0	
4	United States	24.9	
5	Netherlands	24.7	
6	Germany	22.4	
7	Sweden	21.9	
8	Spain	19.9	
9	France	17.6	
10	Portugal	16.2	
11	Belgium	15.6	
12	United Kingdom	13.0	
13	Canada	11.6	
14	Italy	10.2	
15	Brazil	9.2	
16	Japan	8.2	
17	Australia	7.9	
18	Mexico	6.8	
19	Poland	6.3	
20	China	6.2	
21	Colombia	6.1	
22	Russia	4.9	
23	Turkey	2.3	
24	India	0.6	

Figure 1. IoT devices online per 100 in 2015.

In particular, sensors are used many different fields due to the development of sensor technology. Environmental sensing that could use urban planning, electricity, energy management, transportation system and intelligent shopping systems. Biological sensors could use healthcare and medicine systems (Atamli and Martin, 2014).

Table 1. Field of IoT application.

Field of Application	Application
Energy	Smart devices, Energy Management system, Energy Control system
Smart Home	Fire alarm system, Safety control system, Building Automation system
Environmental Monitoring	Air Pollution, Noise Monitoring, Waterways, Industry Monitoring
Green Agriculture	Green Houses, Compost, Irrigation Management, Soil Moisture Management.
Retail & Logistics	Supply Chain Control, Intelligent Shopping Applications, Smart Product Management, Item Tracking, Fleet Tracking
Smart Transportation	vehicular communication, smart traffic control, smart parking, electronic toll collection systems
E-Health	Patient monitoring, Doctor tracking, Personnel tracking, Real-time patient health status monitoring, Home health care

IoT devices have important role due to development of healthcare systems. Healthcare quality is improved with IoT devices such as biological sensors.

In addition, IoT devices are used different fields such as media, production, smart home, smart city, transportation, etc. Table 1 demonstrates field of IoT applications.

1.1 Environmental monitoring

Internet of Things is important for environmental analysis and monitoring in order to ensure environmental protection and control. Sensors are used due to measure the air, water, soil pollution. Internet of Things devices diffuse a large geographic area and collect data from large area.

In addition, Internet of Things applications provide that the checks of use of environmental resources. These applications are used to analysis environmental pollution and make right decision about using of environmental resources (Lee and Lee, 2015).

1.2 Infrastructure management

Urban and rural infrastructure management is important for countries. Therefore, every country has applications due to monitor and control dam, road, bridge, railway tracks, subway, and other critical infrastructure (Jayavardhana et al., 2013). Internet of Things (IoT) devices can be used to monitor and operate events or changes in structural conditions that can compromise safety and increase risk.

Furthermore, Internet of Things (IoT) applications are used for providing access to ships transition from bridge, measure and control the density of road vehicles, check dams' occupancy rate etc. They have big role to coordinate between different service providers and users in critical infrastructure in order to ensure cost effective and time schedule maintenance and repair. Usages of IoT devices provide improving the quality of service like incident management and emergency response (Michael et al., 2014).

1.3 Building, home automation and energy management

Devices has become smart with connect internet. So, systems used in buildings and houses began to be automated. People have become control devices in their home remotely. These smart devices are television, heater, air conditioning, and fridge in the home. Smart buildings include Fire alarm system, Safety control system, Building Automation system, Energy Management system, Energy Control system, Central, Control and monitoring system, Contact communication systems, Centralized information sharing service (Shrouf and Miragliotta, 2015).

Home automation systems, like other building automation systems, are typically used to control lighting, heating, ventilation, air conditioning, appliances, communication systems, entertainment and home security devices to improve convenience, comfort, energy efficiency, and security.

1.4 Transportation

IoT devices are important for transportation. They are used for control transport system. Application of the IoT extends to all aspects of transportation systems. All smart system is communicating each other in order to provide high quality transportation services. Smart transportation services are smart traffic control, smart parking, electronic toll collection systems, logistic and fleet management, vehicle control, and safety and road assistance (Alvi et al., 2015).

1.5 Medical and healthcare systems

The developments of health technology are presenting new opportunities and facilities in order to improve healthcare sectors. In these days, healthcare system use sensors, mobile devices. Internet of Things (IoT) devices are used all parts of healthcare systems.
Figure 2 shows sensors in the healthcare.

They are used for remote health monitoring and emergency notification systems. Monitoring system can enable patient monitoring for chronic issues, checkups, blood pressure and heart rate monitors and unnecessary appointments (Lubecke et al., 2014)

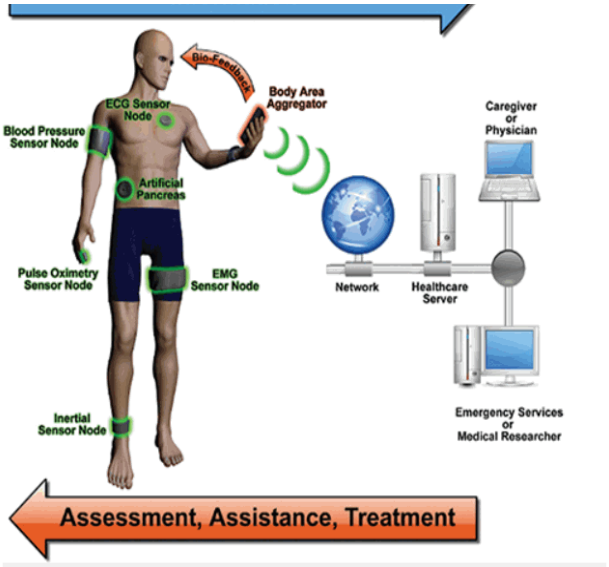


Figure 2. Sensors in the healthcare.

2 Related Work

There are many articles covering security of Internet of Things (IoT) devices in healthcare systems. Nguyen et al. (2015) define the security requirement for highly interconnected network of heterogeneous devices in Internet of Things (IoT). Kai et al. (2013) propound security and privacy methods for Health Internet of Things in order to protect patients' healthcare data. Neisse et al. (2015) propose a Model based Security Toolkit due to the control and protection of user data from Internet of Things (IoT).

Hamdi and Abie (2014) propose a game based model for security eHealth applications in the Internet of Things (IoT). They use security effectiveness and energyefficiency methods for evaluating security strategies.

3 Threats of IoT Health Applications

IoT devices are important for health applications. IoT devices collect measurable and analyzable healthcare data in order to facilitate the work healthcare applications. Therefore, security of IoT healthcare applications is important for healthcare systems. IoT devices are threatened by many security vulnerabilities. I give details about these vulnerabilities in this part of article.

3.1 Energy Optimization

Sensors are significant devices for healthcare systems. Measurable and analyzable healthcare data are gathered very easily with the development of sensor technology. In recent years, wearable devices are very popular for healthcare systems. Wearable devices could collect many healthcare data without disturbing patients. However, energy consuming is important problem for wearable devices. Because wearable devices are small and they are used to collect data form people body. They collect healthcare data from body continuously. Battery is not enough to collect and send health data to healthcare applications. In addition, battery of wearable devices is necessary to be constantly charge. These are serious problems for IoT devices in healthcare systems (Decuir, 2015).

3.2 Privacy

Privacy has many definitions in the literature. Privacy is important topic for information security at healthcare system in the world. Hence, many international organizations define privacy. The Organization for Economic Cooperation and Development (OECD) defines it as "any information relating to an identified or identifiable individual (data subject)" (Chen and Zhao, 2012).

Healthcare data are collected from IoT devices. These devices gather data by remote access mechanisms which have some challenging about privacy and security. Data collected by the sensor is transmitted to the database or cloud over internet. In addition, IoT devices connect internet and communicate with each other from the Internet. Security vulnerabilities on Internet and IoT devices are threatened health data. Additionally, healthcare data are collected from different health units. Health data is shared by various health units. Every unit must provide privacy of data. Because healthcare data includes essential significant information. All the world's attackers all the world's attackers want to capture health data. Thereof, privacy of data must be protected (Thilakanathan et al., 2013).

3.3 Trust

Trust management is important for IoT devices and applications due to provide security and privacy of data. Because all devices connect network and send data to applications. Therefore, devices on the internet must be trusted due to ensure privacy and security. Attackers could connect device IoT applications in order to manipulate data (Skarmeta et al., 2014).

Data collection trust is serious issue because of huge volumes of data are collected from devices. Big data is used by IoT health application owing to make right decision about patients and improve quality of healthcare. Moreover, IoT health care services include data process, analysis and mining. Attackers could be damaged big data with create damage or malicious input of IoT devices. Hence, researchers study about challenges of trust management in IoT. Trust management in IoT must implement network layer and application layer (Abomhara and Køien, 2004).

3.4 Denial-of-service attacks (DoS)

Denial-of-service attacks (DoS) are to make IoT devices and IoT applications cannot provide service. IoT devices connect network and transfer data and communicate with each other. IoT applications need to connect to network and receive data from these devices. Denial-of-service attacks (DoS) dangerous attacks for IoT applications because of machine or network resource. IoT devices have low memory capabilities and limited bandwidth, battery, and disk space. Hence, they are affected from Denial-of-service attacks (DoS) easily (Abomhara and Køien, 2004).

3.5 Physical attacks

Physical security is serious issue for IoT devices because of gathering data from of unprotected

environment. Further, IoT devices are small devices and they are integrated TVs, cars, air conditioning, ovens etc. Therefore, these devices could be stolen easily or changed configured settings. Attackers can change data sent by IoT devices. IoT devices areexposed many physical attacks such as a secret stealing, software manipulation, and hardware tampering.

3.6 Data Manipulation

Data is important IoT healthcare applications. Data is used all steps of healthcare systems. Thus, attacks are against to data security and privacy. Attacks are stealing data, data manipulation and damaging data. Health data is important and sensitive data for all country in the world because of including personally identifiable data. Data is used multiple paths in the IoT. These are IoT devices (data generator, data receiver, and aggregation point), the internet (multi-directional data transport), the cloud (data stored), the machine (application services, big data repositories, analytic) (Bing et al., 2011)

Attackers steal health data for malicious use and they damage victims. They steal data when data is generated and transported by IoT devices. Attackers manipulate or change data in order to redirect victims what attackers want. Doctors could give wrong decision about diagnosis and treatment because of data manipulation. In addition data loss is serious problem for IoT health application.

4 Security Solution of IoT Health Applications

Security is important IoT health applications because of sensitive health data privacy. This section classifies security solution in order to protect data from attacks. Figure 3 represents data security in IoT healthcare applications.

Figure 3. Data security in IoT healthcare applications.

Access Control is important step in protecting IoT healthcare applications and health data. Well-designed access control must be implemented IoT healthcare applications and devices. IoT devices collect health data from patients and transfer health data to healthcare databases. Hence, IoT healthcare applications must have strong access management in order to ensure healthcare data security and privacy. Through access control systems, an organization can restrict andmonitor the use of critical data, and protect privacy and security (Rush Carskadden, 2013).

In addition, employers in healthcare should have the awareness of information security owing to provide security of IoT healthcare applications and health data. Information security awareness training should be given to healthcare staff members. Furthermore, employees should receive all training about access control of priority for ensuring data security, privacy, and patient rights.

Access management in IoT healthcare applications protects to misuse healthcare data and perform malicious attacks on the users' healthcare data.

IoT health devices are tiny and integrated other systems. IoT health devices collect data from various environments. Hence, physical security is significant topic for IoT health devices. IoT health devices should be secure against physical threats (Shen and Liu, 2011).

Physical security of IoT health devices and health data involves protection against environmental threats, accidents, physical sabotage, and theft. IoT health devices should have replacement devices for protecting physical attacks. In this way, IoT health devices keep on collecting and transferring data. Countermeasures should be taken to decrease the damage and recover from attacks, accidents and disaster quickly.

Network security is important issue for IoT health devices and applications. All IoT devices connect to network and communicate each other over network. Therefore, network is required in all steps in the IoT health applications'. Some technologies are WiFi, Bluetooth, Zigbee, RFID etc. Especially, wireless body area network (WBAN) is used to data collect from wearable devices.

Firewalls, IPS, IDS, ingress/egress filtering structures, internet protocol security (IPsec), secure Sockets Layer/Transport Layer Security (SSL/TSL) should be used in order to ensure network security. HTTPS is used for application to encrypt to message (Sadeghi et al., 2015; Xingmei et al., 2013). Data privacy is major issue for IoT health devices and applications because of the ubiquitous character of the IoT environment. IoT health devices connected each other and send data. Strong encryption algorithm is used for data. Data must be encrypted and sent IoT application over a secure network. RFID technology is used for IoT devices to send data. However, RFID has

some security vulnerabilities such as reverse engineering, eavesdropping, man-in-the-middle attack, spoofing etc. When data is sent by RFID technology, some security measures are taken owing to provide privacy protection (Xingmei and Jing, 2013).

Besides, IoT health application should have strong access control management and trust management services due to ensure data security and privacy. Health data is collected from IoT health devices then share various health units. Healthcare data is used accurate assessment and right decision about patient treatment and diagnosis. IoT healthcare application must have "need-to-know" principle for authorization management (Weinberg et al., 2015).

Policies**,** standards and guidelines could be developed, documented, and implemented. Further, many standards about security of healthcare are published by international organizations. These documents are used for on account of providing security and privacy. Each employee and department should have enough information about procedures, guidelines, and standards related to data security and privacy. They could be reviewed and update regularly and change according to the needs of the healthcare sector (Weinberg et al., 2015). Trainings should be prepared owing to improve information security awareness of healthcare staff. These trainings give details about fundamental security and risk IoT healthcare applications and emphasize about privacy of health data. These trainings could be provided to employees regularly (Weinberg et al., 2015).

All IoT healthcare devices, IoT healthcare applications and network components' log must be collected with central log management systems. Logs are monitored, analyzed and evaluated so as to prevent unwanted events to healthcare systems. Besides, central log management or security information and event management (SIEM) must have auditing to ensure security. Undesirable events must be reported security team quickly to interfere unwanted events. Central log management or security information and event management (SIEM) must have strong authentication and authorization to monitor the audit log. The log should be checked continuously.

Unfortunately, auditing is a passive defense because of becoming aware of critical security event after the occurrence of the event. Auditing help people to response to unwanted-event quickly.

5 Conclusions

IoT devices are very important for systems. Today, many systems have become smart with IoT devices. These systems include big data. IoT devices collect data for these systems. Data is sensitive because of including personal information. Hence, these systems have many threats about data security and privacy. Security recommendation could be used to mitigate the security threats.

This paper presents detail about IoT healthcare applications and security threats in IoT application. In addition, this paper gives security solution in order to mitigate security threats.

References

M. Abomhara and G. M. Køien. Security and Privacy in the Internet of Things: Current Status and Open Issues, 2004.

S. A. Alvi, B. Afzal, G. A. Shah, L. Atzori, and W. Mahmood. Internet of multimedia things: Vision and challenges, 2015.

A. W. Atamli and A. Martin, Threat-based Security Analysis for the Internet of Things, IEEE, 2014.

C. Bing, D. Yuebo, J. Bo, Z. Xiang, and Z. Lijuan. The RFID-based Electronic Identity Security Platform of the Internet of Things, 2011 International Conference on Mechatronic Science, Electric Engineering and Computer Jilin, China, August 19-22, 2011.

D. Chen and H. Zhao. Data security and privacy protection issues in cloud computing, International conference on computer science and electronics, engineering, 2012.

J. Decuir. The Story of the Internet of Things, IEEE Consumer Electronics Magazine, 2015.

M. Hamdi and H. Abie. Game-Based Adaptive Security in the Internet of Things for eHealth, IEEE, Communication and Information Systems Security Symposium, 2014.

G. Jayavardhana, B.Rajkumar, M. Slaven, and P. Marimuthu. Internet of Things (IoT): A vision, architectural elements, and future directions. Future Generation Computer Systems, 2013.

K. Kai, P. Zhi-bo, and W. Cong. Security and privacy mechanism for healthinternet of things, 20(Suppl. 2): 64–68, www.sciencedirect.com/science/journal/10058885 http://jcupt.xsw.bupt.cn, December 2013.

I. Lee and K. Lee. The Internet of Things (IoT): Applications, investments, and challenges for enterprises, Elsevier, 2015.

O. B. Lubecke, X. Gao, E. Yavari, M. Baboli, A. Singh, and V. M. Lubecke. E-Healthcare: Remote Monitoring, Privacy, and Security, 2014, IEEE, page 1351-1360, 2014.

C. Michael, L. Markus, and R. Roger. The Internet of Things, McKinsey Quarterly, McKinsey & Company, Retrieved 10 July 2014.

R. Neisse, G. Steri, I. N. Fovino, and G. Baldini. SecKit: A Model-based Security Toolkit for the Internet of Things 2015.

K. T. Nguyen, M. Laurent, and N. Oualha. Survey on secure communication protocols for the Internet of Things, 2015.

R. Sadeghi, C. Wachsmann., and M. Waidner, Security and Privacy Challenges In Industrial Internet of Things, 2015.

A. Santos, J. Macedo, A. Costa, and M. J. Nicolau. Internet of Things and Smart Objects for M-Health Monitoring and Control, CENTERIS 2014 - Conference on ENTERprise Information Systems / ProjMAN 2014 – International Conference on Project MANagement / HCIST 2014 - International Conference on Health and Social Care Information Systems and Technologies, 2014.

Michael J. Rush Carskadden, Threat Implications of the Internet of Things, 5th International Conference on Cyber Conflict, Tallinn, 2013.

M. Shane, V. Nicola, S. Martin, and L. Anne. The Internet of Everything for Cities: Connecting People, Process, Data, and Things To Improve the 'Livability' of Cities and Communities, Cisco Systems, 2014.

G. Shen and B. Liu. The visions, technologies, applications and security issues Of Internet of Things, 978-1-4244-8694-6/11/, IEEE, 2011.

F. Shrouf and G. Miragliotta. Energy management based on Internet of Things: practices and framework for adoption in production management, Journal of Cleaner Production, 2015.

S. Sicari, A. Rizzardi, L.A. Grieco, and A. Coen-Porisini. Security, privacy and trust in Internet of Things: The road ahead, Computer Networks, 2015.

A. F. Skarmeta, J. L. Hernández-Ramos, and M. V. Moreno. A decentralized approach for Security and Privacy challenges in the Internet of Things, IEEE World Forum on Internet of Things (WF-IoT), 2014.

D. Thilakanathan, S. Chen, S. Nepal and R. and A. Calvo. Secure and controlled sharing of data in distributed computing, IEEE 16th International Conference on Computational Science and Engineering, 2013.

R. H. Weber., Internet of things – new security and privacychallenges, Computer Law & Security Review, 2010.

B. D. Weinberg, G. R. Milne, Y. G. Andonova, and F. M.Hajjat. Internet of Things: Convenience vs. privacy and secrecy, Business Horizons, 2015.

International Telecommunication Union. ITU Internet Reports, the Internet of Things, 2005.

http://www.ibmbigdatahub.com/blog/privacy-and-internet-things, access date: 27.01..2016.

http://www.iso.org/iso/home/search.htm?qt=health+privacy&sort= , access date: 27.01..2016.

X. Xiaohui, Study on Security Problems and Key Technologies of the Internet of Things, 2012.

X. Xingmei, Z. Jing, and W. He, Research on the Basic Characteristics, the Key Technologies, the Network Architecture and Security Problems of the Internet of Things, 2013 3rd International Conference on Computer Science and Network Technology, IEEE, 2013.

Simulation of Data Communication System taking into Account Dynamic Properties

Galina M. Antonova Vadim V. Makarov

The faculty of Cybernetics, National Research Nuclear University MEPhI, Moscow, Russia, https://mephi.ru/
Trapeznikov Institute of Control Sciences ICS RAS, Russian Academy of Sciences, Moscow, Russia,
gmant@ipu.ru, makfone@ipu.ru

Abstract

This paper continues the study, presented at the 8th EUROSIM Congress on Modeling and Simulation and devoted to creation of algorithm and simulation model of network functioning, taking into account dynamic characteristics of the network in condition of variable relationship signal-to-noise. Simulation algorithm was augmented for adequately representation of the state of the real network i.e. possible changes of topology due to the link failures and disabling individual nodes. It is possible to expand the capabilities of the model presented in the 8th Congress as a simulation model of Information Flow on Transport Layer of Open System Interconnection Model. The current version of the model realizes input of the adjacency matrix describing the network topology, the algorithm of the path search by Dijkstra on the network level, and simulation of the loss of connection. So the main goal of the new paper is to bring the structure of the model to the structure of the real network and to check the possibility of transferring a given amount of information in conditions of interference by means of evaluation of coefficient of readiness for Data Communication System.

Keywords: modeling, Monte-Carlo simulation, information technologies, algorithm

1 Introduction

Modeling of the dynamic properties of the Data Communication network is one of the urgent tasks in the modern theory of communication. Processes of data transmission have very high speed and the data transfer equipment includes specialized devices that provide perform the necessary calculations at speeds far exceeding the speed of imitation of processes of data transmission in modern software environments. At the same time, quite often in practice there is a need to evaluate the number of criteria of quality that are related not only internal properties of hardware and data transmission protocols but external integral characteristics.

They are related to external characteristics of the ne twork as a whole or its separate fragments. These characteristics usually are interested for the creators of information systems. One of the main criteria of quality is the availability of Data Communication System (DCS). The other important criterion is the probability of message delivery in the presence of noise. These criteria help to check the properties of the fragment of network or the network as a whole and test its suitability for the solution of the problem of the transfer of large amounts of information at any given time.

The throughput for different channels of network and communications centers and the procedures of changing network topology must be realized as part of the overall procedure of simulation of DCS for evaluating dataflow and network capacity in system as a whole. Thus the paper deals with a problem of construction of adequate simulation model of DCS taking into account dynamic processes in separate channels and nodes of system as a whole. Usually dynamic processes are emerged in form of various violations of the network topology, i.e. in node failures, in communication failures in the communication channels, in the packet loss and in the receiving an increased number of errors due to the increase of the level of interference in communication channels. A universal tool must be created for research and testing of various algorithms of network operation and for correction of new protocols at the stage of pre-project inspection.

The structure of paper is as follows. Introduction demonstrates some distinctive features of the statement of the problem. The section 2 is devoted to describing of a static variant of the simulation model. It contains brief analysis of results of imitation experiments. In the section 3 algorithm of simulation of DCS with variable topology and quick details about modeling computer's program are described. The section 4 involves simulation results for checking proposed algorithm.

For a formal definition of DCS designed for the transmission of big volume of information represented in a form of graph structure, we can use the definition in the form of a functional signed graph from (Shul'tsc, Kulba *et al*, 2011). A tuple of parametric functional graph $<(X, E), V, W, U>$ is introduced. In this recording

$G = (X, E)$ – directed graph, describing the structure of DCS, symbol V – a set of parameters of the vertices of the graph, where every vertex is according to node of network. Parameters of the vertices must define all characteristics of nodes necessary for modeling of DCS and evaluating its' criteria of quality. For example, parameters define procedure of modeling host-router, involved in the selection of the direction of information transmission, or node – repeater designed only to enhance the signal to increase the communication distance. $V = \{Pnode(x), x \in X\}$, where $Pnode(x)$ – a set of parameters of the vertex (node) x. Symbol U describes the parameter space, a W – weights of edges, simulating communication channels. Weights of edges can change during modeling, and finding the path for message transfer.

The main directions of modeling of DCS in modern theory of communication involve three different approaches.

First of all modeling of the propagation medium of the signal allows investigating new principles of design of equipment for communication. Mathematical and statistical models help to create estimates of main characteristics of communication technique for data transferring. They help to define requirements to new equipment. Such models are described in (Proakis, 1995; Rappaport, 2002; Saleh, Valenzuela, 1987; Spencer, Rice *et al*, 1997) and they expand possibilities of designing of new DCS.

Secondly the modeling of the functioning of DCS as complicate technical system opens new directions for optimization and improvement of existing constructions of DCS. A huge number of models are created over the years of development of the theory of communication. These include works (Foschini, 1996; Antonova, 2007), devoted to data transmission in condition of fading, and many other publications. For example, a set of works are the well-known (Irvine, Harle, 2001; Tanenbaum, 1996) and others.

Thirdly special direction is devoted to different modern problems. At the moment models for describing of various indicators of quality of service (QoS) have been actively developed. For example, for the transfer of information via packets the set of basic criteria involve Bandwidth, Delay, Packet loss and Jitter. A significant number of modern publications including the Russian-speaking illuminate this direction.

2 Algorithm of Simulation of Data Communication System

2.1 Statement of Simulation Problem

The problem of modeling the networks' dynamic properties may be solved by means of imitation statistical model of DCS, proposed in (Antonova, 2013; Antonova, Titov, 2011; Antonova, Kolutcsky, 2015). DCS is represented as a set of channels and communication centers. Large centers for data gathering create a set of signal office centers. In common with channels for information transferring they create graphical image of network topology with variable structure. In the existing networks information may be transmitted in different directions. So almost every node can be both a starting and an end node in the procedure of message delivery to the addressee. This fact determines dynamic changes in the network.

In order to simplify debugging and simulation procedures the assumption is introduced that data is transmitted in one direction from one initial vertex to the destination vertex in network fragment under consideration. This assumption simplifies the development of simulation algorithms, but it is a limitation in the development of a universal instrument for studying processes of transferring large amounts of data in DCS. However imitation statistical model will allow checking quality of data transferring by means of imitation experiments for dynamic stochastic DCS according to algorithm from (Antonova, 2013; Antonova, Titov, 2011; Antonova, Kolutcsky, 2015).

The proposed statement of the problem for simulation of dynamical features of DCS is considered in detail in (Antonova, Titov, 2011). The first variant of a simulation model for the static mode of operation and single state of network topology is presented in flow chart of simulation algorithm in Figure 1. It does not use the possibilities of modern software tools and focused on the tedious procedure of the source data preparation for modeling. Symbol S represents the number of message. Symbol Sk is labeling maximal quantity of messages under simulation.

Matrix $MK(IJ)$ contains symbol 1 on position with number of row I (node with number I) and column with number J (channel with number J) if such channel exists in network topology. Symbol $Jmax$ defines maximal quantity of channels, outgoing from current node with number I in accordance with network topology. Symbol Ik is maximal quantity of nodes.

Symbol Jt is a number of channel selected in simulation procedure according to special criterion. This channel connects the current node with node $I(Jt)$ in accordance with network topology.

After start of algorithm the necessary input data is read from a file. The movement of the first message from the first node to next node according the topology of fragment of network under consideration begins with the scanning of outgoing channels. All outgoing channels are described in matrix $MK(IJ)$. This matrix is continuously visible in the search of the outgoing path.

Simulation for every outgoing channel is fulfilled for transfer time evaluation. The level of errors for given signal-to-noise ratio or for given imitation model of transfer process is also determined. The channel is selected to send a message when the cycle of viewing

for channels leaving the current vertex is over. Selection criteria may be different.

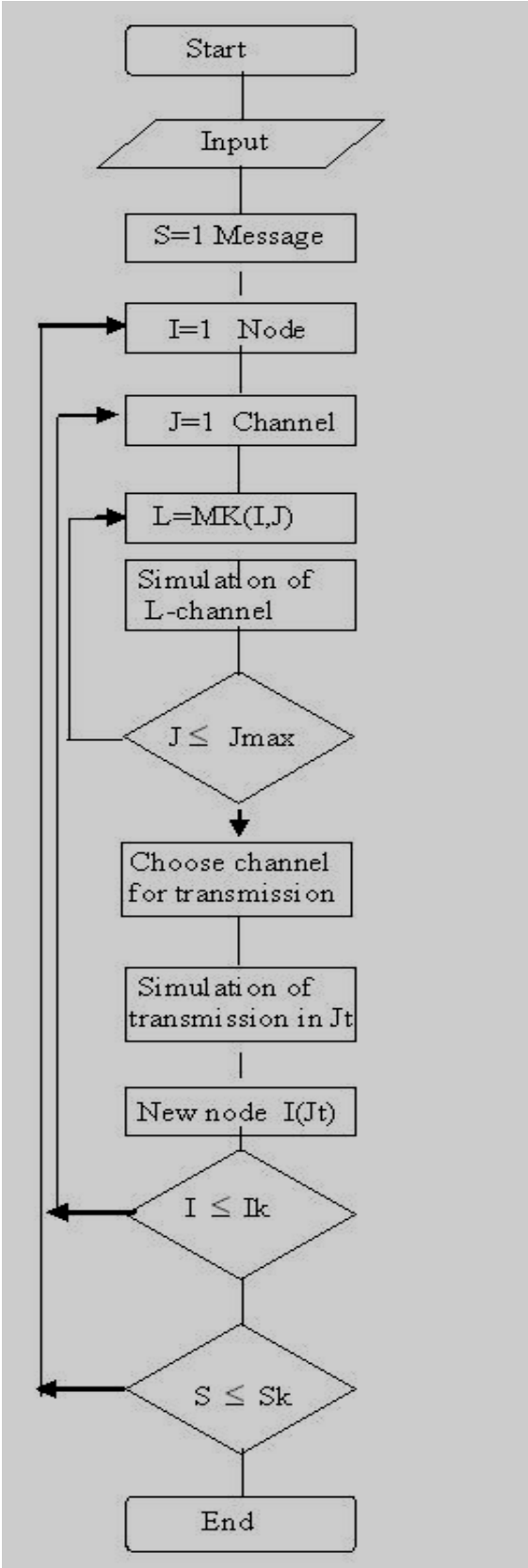


Figure 1. Flow Chart of a static variant of the simulation model.

After the end of the simulation of the transfer process new node is fixed. It will be that vertex of the graph in which the movement of the first message will stop. For new node all procedures for choice and simulation of the transfer process in the channel will be repeated. Thus the first message will continues the moving to final vertex of the graph. When message arrives in final vertex *Ik,* the number of message is tested. If transferring of information didn't over the next message is selected.

Further the moving of the following message will be simulated since first node and so on. At the end of list of messages when last message with number *Sk* will arrive in final node the availability coefficient will be calculated by means of definition of ratio of time of useful moving of messages along fragment of network to common time of simulation of the transfer process. This value is saved in special variables.

This scheme will be supplemented with the outer loop managing random changes in network topology under conditions of noise in functioning of communication channels and nodes. Such model will reflect the variable structure of the computer network in the DCS, i.e. the possibility of link failure or the blocking nodes of a communication network, and subsequent recovery as a result of urgent repair. It will allow checking the impact of interference on the delivery time of messages in the network. This model allows to estimate the availability of the network and to provide additional opportunities for evaluation of the possibility of Big Data accumulation.

2.2 The first results of modeling

The first variant of modeling program was realized by means of Pl-language and ES 1045 computer. It allowed evaluating transmission characteristics of the information flow for static mode of operation of the network under interference. This model did not take into account the variable structure of the network under real conditions of information transferring but was used in imitation experiments for testing and evaluating networks characteristics in noise conditions (Antonova, 1996; Antonova, 1999; and Antonova, 2007).

3 Simulation of Data Communication System with Variable Topology

The considered variant of simulation model of DCS was raw realization of proposed simulation algorithm by means of C++ language and personal computer. It created the basis for detailed implementation of the model reproducing the dynamic structure of data transferring network (Antonova, 2013; Antonova, Titov, 2011; Antonova, Kolutcsky, 2015). The flow chart of a new simulation algorithm is shown in Figure 2. Additional symbol *Nt* in this figure defines the

number of variant of network topology under consideration in simulation procedure. New variant of topology may be appearing only for new message. It limited models possibilities but it was necessary for testing simulation algorithm. New topology may involves the channel interference, the communication gap between individual nodes and the shutdown of the node, reducing noise, the restoration of the link between nodes, the repair of nodes.

A new variant of the simulation model is implemented in the environment of Microsoft Visual Studio 13. The developed algorithm simulates the process of transferring messages within the network fragment which is represented as oriented weighted graph. The vertices of the graph, i.e. the network nodes, simulated work of the packet switches via the TCP/IP Protocol with realization of procedures of determining the optimum route of messages transmission.

Edges of the graph simulate the communication channels with a certain bandwidth determined by the current value of the ratio signal-to-noise. In input data for simulation model the network graph is interpreted as the adjacency matrix, each element of which records the presence or absence of communication between nodes in network.

For adaptation to possible changes in the real network such as communication gap, a node failure, and addition of the node with specified links, an adjacency matrix reflecting the network topology duplicates the matrix allowing for program user to dynamically track the topology changes in the process of simulation program functioning.

Weights of connections between network nodes are defined by the relative time of messages' delivery between nodes according to the ratio, which varies with the changes in noise power:

$$T = 1/C, \quad (1)$$

where C – throughput capacity of communication lines that are installed dynamically from the known ratio of the C. Shannon:

$$C = F\,log(1 + P_S/P_{Sh}), \quad (2)$$

where F - is the bandwidth of the communication line, P_S – signal power, P_{Sh} - noise power.

A minimal message delivery time is selected as a criterion for choice an optimal route for messages transmission between the start and end nodes of the network in the simulation model.

Messages are transferred between the specified fixed network nodes (start and end nodes). At each vertex for determination the channel for transmission the optimal route to the destination node is searched and for transmission the first link of the found route is selected.

For the next point of the route the optimal route to the destination network node is searched again and so on until the end node will be reached. Finding of the optimal route based on the algorithm, which is used famous Dijkstra's algorithm.

Each message from the stream in the process of the simulation model functioning appears in the window indicating the route to the destination node and the relative time of motion.

The following objects are defined in proposed algorithm: the array of distances between vertices; active vertex; dynamic list of available vertices from the active vertex (the list of visited vertices). The dimension of the array of distances corresponds to the rank of the adjacency matrix. Each element of the distances array is a structure. It includes the distance from the initial vertex to the vertex corresponding to the element of the array with current number (the initial distance is infinity) and the list of vertices making the route from the initial vertex to the considered vertex from the array (at the initial moment the list is empty).

The active vertex is special variable. It contains number of the adjacency matrix row when the search algorithm fulfills definition of outgoing channels. The list of visited vertices is dynamical array. According algorithm vertices connected with active vertex are added in that list and visited vertices are deleted. At the beginning of the algorithm initial node is an active vertex. The list of vertices making the route for first element of the array of distances will contain initial node and the array of distances from initial node will contain zero. The optimal route search procedure involves following stages: the adjacency matrix row associated with active vertex is examined; adjacent vertices are defined and added in the end of the list of visited vertices. The node having visited vertices is deleted from the top of the list. The node, which became the first in the list, converts to active vertex. The distance from initial vertex to vertex associated with active vertex is defined. If it is less than distance fixed in the array of distances early, than both distance and route in the array of distances are corrected.

The algorithm ends when the list of visited vertices is over. The criterion for the optimal route search is value of the message delivery time. The optimal route is defined with using matrix contained values reverse of throughput of networks channels. These values are created automatically in procedure of algorithm realization. The relationship

$$Tp = 1/log(1 + P_S/P_{Sh}), \quad (3)$$

defines value of each existing connection.

Value Tp is formed by means of simulation of the signal-to-noise relationship. Different distribution laws known from the results of statistical studies of DCS may be used for describing the signal-to-noise relationship. The average value of the signal-to-noise relationship is chosen equal to 100. The system timer controls changes in topology. It is the lack of simulation model.

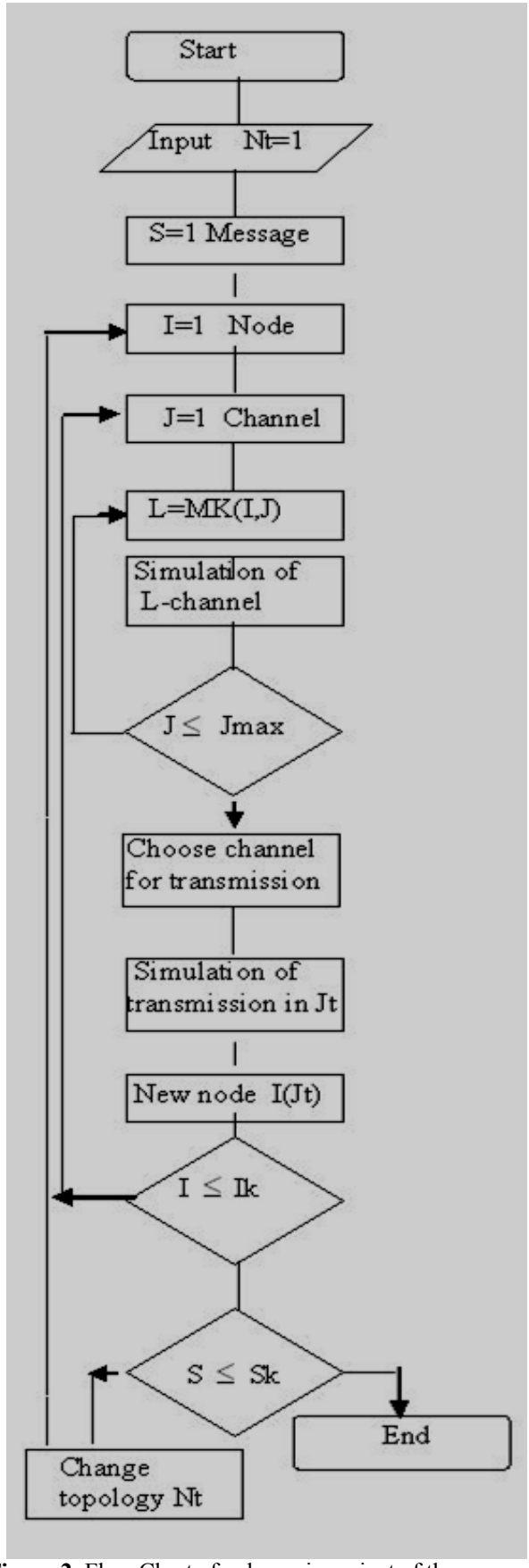


Figure 2. Flow Chart of a dynamic variant of the simulation model.

Every message from information stream appears in the window of simulation model. The route to final node and time of motion are demonstrated in window. All modeling events are fixed in track of events according to order of their receipt.

All changes to the network structure dynamically adjusted in the adjacency matrix. Because of calls to the adjacency matrix are happen from different threads in a model the mutual exclusion of these flows is organized. To do this a synchronization object "mutex" is used. This object is installed in a special signaling state if not busy by any thread. This object at any point in time can only hold one thread that prevents simultaneous access to a shared resource.

4 Results of Simulation Experiments

Initial network topology is selected in the form shown in the Figure 3. The adjacency matrix *M* describing this network fragment will sets by user of program in special Windows Forms with the use of element of control Data Grid View.

$$M = \begin{pmatrix} 1&1&1&0&0&0&0&0&0&0 \\ 0&0&0&1&1&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&0&1&1&0&0 \\ 0&0&0&0&0&0&0&1&1&0 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&1 \end{pmatrix}.$$

Weights of connections are established dynamically in the process of the functioning of simulation model. According to simulation algorithm experiment consists in transmission of message flow from the initial vertex of the network fragment to a final vertex of the network fragment. Channel interference, the link failure and disabling individual nodes are considered.

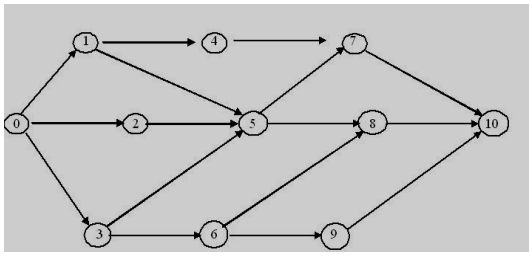


Figure 3. Scheme of a fragment of the communication network.

Results of simulation experiment are shown in Table I. The transmission of 30 messages in condition of change relationship signal-to-noise, the link failures and disabling individual nodes are considered. Table 1 contains results from Windows Forms: number of message (first column); time of message transmission

(second column); the order of the passage of nodes from network fragment (third column); delivery information about messages (fourth column). Because of a long period of disconnection in the network fragment is simulated, the last three messages are not delivered.

5 Conclusions

A series of simulation results is only a checking of proposed algorithm. It needs in quality programming and extensive testing in terms of the input variable factors reflecting the structure of real networks. Big volume of initial input data are needed for creation of detailed and adequate estimates of performance and availability of DCS, but this problem is much easier then the problem of creating models of dynamic network for data transfer. The instruments for control of information flow in real network involve the repair of failed channels, addition of new channels, increasing of velocity of data transmission and extension of channel bandwidth by means of installation of new equipment and others. Further improvement of the model consists in the development of the structure of the program modules for simulation of different types of communication channels, support modules and the control program, taking into account the received results. Once the application package has been validated, the adequate results of simulation may forecast a situation in functioning of equipment of communication network under conditions of large load.

Table 1. Simulation results.

Message	Characteristics of the motion		
	Time	*Route*	*Path of events*
1	1.015790	0 1 4 7 10	Start End
2	1.226979	0 1 4 7 10	Start End
3	1.101203	0 2 5 8 10	Start Channel 8-10 End
4	1.212878	0 3 6 9 10	Start End
5	1.218719	0 3 5 7 10	Start Node 6 End
6	1.197147	0 2 5 7 10	Start End
7	0.992137	0 1 4 7 10	Start Channel 3-5 End
8	1.211201	0 1 5 7 10	Start End
9	0.998856	0 1 5 7 10	Start End
10	1.128747	0 1 5 7 10	Start Node 9 End
11	1.116496	0 1 4 7 10	Start Channel 1-5 End
12	1.170238	0 2 5 7 10	Start End
13	1.124116	0 1 4 7 10	Start End
14	1.362350	0 1 4 7 10	Start End
15	1.245929	0 2 5 7 10	Start Channel 0-2 End
16	1.391276	0 1 4 7 10	Start Node 2 End
17	1.212394	0 1 4 7 10	Start End
18	1.364307	0 1 4 7 10	Start End
19	1.302780	0 1 4 7 10	Start Channel 5-8 End
20	1.294338	0 1 4 7 10	Start End
21	1.138025	0 1 4 7 10	Start End
22	1.382752	0 1 4 7 10	Start End
23	1.097262	0 1 4 7 10	Start Channel 5-7 End
24	1.273273	0 1 4 7 10	Start End
25	1.382320	0 1 4 7 10	Start End
26	1.104619	0 1 4 7 10	Start Channel 1-4 End
27			Start A message is lost
28			Start A message is lost
29			Start A message is lost

Acknowledgements

Authors thank lecturer of Moscow Financial-Juridical University K. N. Kolutcsky for conscientious work as a programmer.

References

Galina M. Antonova. Realization of the Optimization - Simulation Approach in the Selection of an Algorithm for the Functioning of Data Communication Systems. *Automation and Remote Control*, 57(9):1357-1363, 1996.

Galina M. Antonova. Choice of Noise Immune Correcting Codes by the LP_τ-optimization within the Framework of the Optimization- Simulation Approach. *Automation and Remote Control*, 60(9): 1347-1352, 1999.

Galina M. Antonova. *The mesh methods of uniform probe for investigation and optimization of the dynamical stochastic systems*. Moscow: Phizmatlit. 2007.

Galina M. Antonova and A. P. Titov. Simulation of information flow in e-Governance. *In Proc. 5-th All-Russian Science - practical Conference on Simulation and its Application in Science and Industry*, St. Petersburg: OAS Center of Design and Shipbuilding. 5(1):325-328, 2011.

Galina M. Antonova. Simulation of Information Flow on Transport Layer of Open System Interconnection-Model. *In Proc. of 8th EUROSIM Congress on Modelling and Simulation,* Cardiff, Wales, UK: IEEE Press, Sep. 2013, pages 567-572.

Galina M. Antonova and Konsyantin N. Kolutcsky. Simulation model of the data transfer process in the network segment. *In Proc. 2-th International Science - practical Conference on "Modern information technologies in professional activity"*, Moscow: MFUA, 2: 15-18, 2015.

J. Irvine, D. Harle. *Data Communications and Networks: An Engineering Approach*. England: John Wiley&Sons. 2001.

G. J. Foschini. Layered space-time architecture for wireless communication in a fading environment when using multiple antennas. *Bell Labs Technical Journal*, 1(2): 41-59, 1996.

J. G. Proakis. *Digital Communications*. NY: McGraw Hill. 1995.

T. Rappaport. *Wireless communications: principles and practice*. New Jersey: Prentice Hall PTR. 2002.

A. Saleh, R. Valenzuela. A Statistical Model for Indoor Multipath Propagation. *IEEE Journal on Selected Areas in Communications*, 5(2): 128–137, 1987.

V. L. Shul'tsc, V. V. Kulba and others, *Information control in condition of active confrontation: models and methods*. Moscow: Science. 2011.

Q. Spencer, M. Rice, B. Jeffs, M. Jensen. A statistical model for angle of arrival in indoor multipath propagation, *IEEE Vehicular Technology Conference*. 47: 1415–1419, 1997.

A. S. Tanenbaum. *Computer networks*. Upper Saddle River, NJ: Prentice Hall. 1996.

Simulation of HTTP-based Services Over LTE for QoE Estimation

Alessandro Vizzarri[1] Fabrizio Davide[2]

[1]Department of Enterprise Engineering, University of Rome Tor Vergata, Italy,
alessandro.vizzarri@uniroma2.it

[2]Department of Innovation and Information Engineering, Guglielmo Marconi University, Italy,
f.davide@unimarconi.it

Abstract

Long Term Evolution (LTE) enables bandwidth consuming HTTP applications as video streaming. Mobile Network Operator (MNO) is committed to guarantee acceptable levels of Quality of Service (QoS) and Quality of Experience (QoE) perceived by the end user. A correlation between the transport informations with the application informations is an important approach to be adopted by the MNO. This correlation is more useful if a second entity, as the Over The Top (OTT), cooperates for the content delivery process. In the scientific literature different mathematical models are used in order to correlate QoE to the QoS. This paper aims at analyse them in case of of HTTP based Web services as HTTP web browsing and HTTP video streaming. Different scenarios are simulated using OPNET simulation software tool. They can differ if the service is fully managed by the MNO (MNO-managed class) or if OTT cooperates with own content (OTT-managed). This is the case of YouTube. Results are analysed through regression k- means clustering techniques.

Keywords: LTE, QoS, QoE, over the top; YouTube; video streaming; key performance indicators

1 Introduction

In the last years, telecommunication technologies are enabling the delivery of bandwidth-consuming applications as web browsing or video streaming of several multimedia objects (Ericsson, 2008). The fourth generation of mobile networks, known as Long Term Evolution (LTE), is the first 3GPP cellular fully-IP standard. LTE is the most advanced technology to satisfy the increasing demand for mobile broadband services. It is able to offer to end users a download data rate up to 100 Mbps and an upload data rate up to 50 Mbps (3GPP, 2007). LTE is also characterized by a flexible and interoperable fully-IP network architecture. We have also a direct management of Quality of Service (QoS) policies based on bearers and QoS Class Identifier (QCI). These informations are managed by Mobile Network Operators (MNOs) in order to efficiently deliver acceptable service levels to the end-users. QoS policies in LTE are mainly focused on measurable parameters called Key Performance Indicators (KPIs), namely bandwidth, delay, jitter, packet loss rate, data rate, priority. These QoS native features are crucial for an efficient network management of both data and voice services. Here we introduce an exercise of correlating the Quality of Experience (QoE) as perceived by the end user to the QoS as measured by the MNO at the network level for the HTTP web browsing and video streaming applications. We will introduce some relevant study cases, grouped in either MNO-managed class or OTT-managed class. We will review related works in Section II and we propose our approach in Section III. Section IV describes the study cases and present results from extensive simulations. Section V presents the mathematical models for the data which well describe in our cases the QoE vs QoS correlation. A final discussion states the application range of the proposed approach and its future improvements.

2 Related Works

Several mathematical models are proposed for QoE vs QoS correlation (M. Alreshoodi, et al, 2013). (H.G. Msakni et alii, 2013) presents the concept of Quality of Service (QoS) and Quality of Experience (QoE) applied to video quality assessment. (A. Vizzarri et al, 2013) present a review of studies on QoS in LTE networks. The Quality assessment methodology are essentially three: subjective, objective and network-based. In the first category, some authors introduce Mean Opinion Score (MOS) as a synthetic indicator of QoE, while network KPIs are assumed as indicators of QoS (ITU-T, 2003). (A. Vizzarri, 2014) analyzes the relationship between QoE in terms of MOS and QoS KPIs in case of Voice Over LTE (VoLTE) service. The objective assessment methodology focuses on the measurement of the signal as it would be perceived by an end user. Objective methods can be divided in three main groups: Full Reference (FR), Reduced Reference (RR) and No Reference (NR) (B. Wang et al, 2009). The FR method is based on the estimation of difference between the source video and the received video. The RR method analyzes a portion of informations extracted from the original video. The NR method predicts the video

quality of the received video without accessing to the source video. The network assessment methodology gives an estimation of QoE on the basis of several QoS KPIs measured at network level. Network KPIs are essentially delay, jitter and packet loss rate (M. Siller, et al, 2003). In (M. Alreshoodi, 2013) a fuzzy-logic approach is proposed. (T.H Truong, 2012) analyzes the relationship between QoE and QoS in case of IPTV applications. Well known papers propose to express the QoE/QoS correlation through statistical analysis. The regression models more frequently used are: logarithmic (Weber-Fechner Law), exponential (IQX Hypothesis) and polynomial (S. Khorsandroo et al, 2012), M. Fiedler et al, 2010). In (J.Urrea , 2015) a multivariate statistical analysis has revealed the factors that influence the quality of a video streaming service on a Multi-hop Wireless Network. Relationships between the QoE and the QoS metrics have been identified via Multivariable Linear Regression (MLR) modeling. This methodology has supported the selection of the relevant factors and the reduction of the model dimensionality. A k-means clustering process allowed the authors to identify some operation ranges in terms of quality that relate the performances to KPIs.

3 Proposed Approach

3.1 Methodology

The methodology adopted for mapping QoE vs QoS in case of HTTP web browsing / video streaming is shown in Figure 1. It is to be remarked that we assume (coherently with some authors and differently from others):

- QoE metrics as Page Response Time (PRT), Video Response Time (VRT);
- QoS KPIs as end-to-end Delay (e2eDEL), Packet Loss Rate (PLR) and Throughput (THR).

3.2 Network models

The considered LTE network models are shown in Figure 2. The model is derived from LTE reference architecture and the related network nodes are the same as defined by 3GPP standard (3GPP, 2007), (3GPP, 2009). HTTP Web/Video Server is the content server containing both HMTL web pages and videos to deliver.

Depending on the owner of the contents to deliver (MNO or OTT), we distinguish for the sake of analysis two management cases for the HTTP web browsing and video streaming services: MNO-managed (class 1) and OTT-managed (class 2). In the first situation MNO is responsible for content ownership, service deliver and service transport. In the second situation MNO is only responsible for transport, while OTT is the content's owner and cooperates with MNO for service delivery. The two classes are shown in Figure 2.

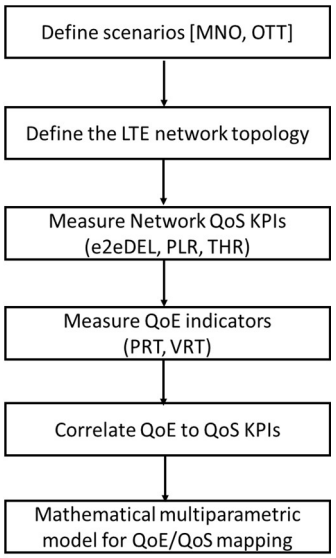


Figure 1. Workflow adopted for QoE/QoS correlation.

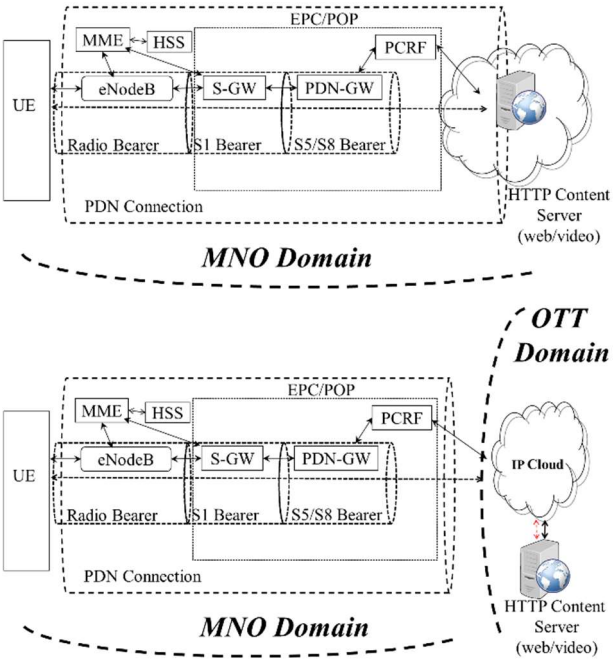


Figure 2. Workflow adopted for QoE/QoS correlation LTE service management classes: MNO-managed (uppermost), OTT-managed (lowermost).

4 Simulation

Network simulations have been performed using OPNET Modeler 17.5 PL6 software tool.

4.1 OPNET Settings

The User Equipment (UE) is modeled with -1 dBi Antenna Gain and -200 dBm receiver sensitivity. eNodeB bandwidth is 20 MHz with 3 antenna sectors; Frequency Division Duplexing (FDD) is the Duplex mode, antenna gain of 15 dBi. As transmission model

we chose a free space model. Evolved Packet System (EPS) bearer is characterized by a QCI equal to 6 (non-Guaranteed Bit Rate) with an Allocation Retention Priority (ARP) equal to 6. Simulation period is 3 minutes.

4.2 Assumptions on users and networks

On the end-user's side we have made a number of reasonable assumptions, as shown in Figure 3. The user sends a first request for a web page in HTML format. After 2 seconds starting from a download of complete web page, user sends a second request for a short video with duration of 20 seconds and a size around 1 MB. The HTML page is composed of: text (180 KB); a figure as top banner (5 KB); 30 small images (7 KB each; 15 large images (20 KB each); a large image (35 KB) as the initial video frame.

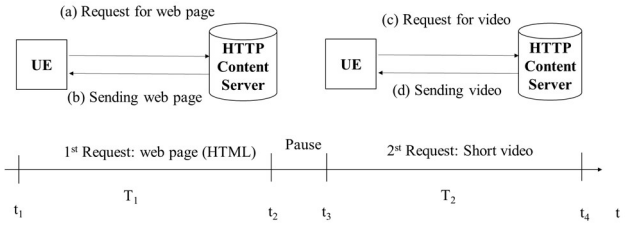


Figure 3. End-user's simulated activity.

On the network side we have modeled the two network models. In the network model of the first class the LTE network topology follows the following criteria:

- unique eNodeB and UE are located in the Rome area;
- EPC and HTTP server, managed by MNO, are located in the Milan area;
- Trunk link type is 1 Gbit/s.

In the network model of the second class (OTT-managed service), the LTE network topology is similar to the previous one but with some differences:

- HTTP server is a YouTube Primary Cache Server located in the Milan area. It is fully managed by You Tube, that plays the role of the OTT;
- The distance between the EPC and the HTTP web server is 100 km;
- An IP cloud is considered, with additive 0.5% Packet Discard Ratio and 50ms Packet Latency. That to take into account impairments due to interconnection between the networks of MNO and OTT.

4.3 Scenarios

Fifteen study cases, called scenarios, are simulated for each class. Table 1 counts 30 scenarios, fifteen cases per class. Scenarios for class 2 are marked with (*).

Table 1. Scenario Configurations – both MNO and OTT Management.

Scenario no.	*Features*		
	UE no.	*S1 External Traffic Load*	*Transmission bitrate over S1[bps]*
1,2,3, 16(*),17(*),18(*)	1	0%; 50%; 75%	1,073,741,824; 536,870,912; 268,435,456
4,5,6, 19(*),20(*),21(*)	5	0%; 50%; 75%	1,073,741,824; 536,870,912; 268,435,456
7,8,9, 22(*),23(*),24(*)	10	0%; 50%; 75%	1,073,741,824; 536,870,912; 268,435,456
10,11,12, 25(*),26(*),27(*)	30	0%; 50%; 75%	1,073,741,824; 536,870,912; 268,435,456
13,14,15, 28(*),29(*),30(*)	50	0%; 50%; 75%	1,073,741,824; 536,870,912; 268,435,456
1,2,3, 16(*),17(*),18(*)	1	0%; 50%; 75%	1,073,741,824; 536,870,912; 268,435,456
4,5,6, 19(*),20(*),21(*)	5	0%; 50%; 75%	1,073,741,824; 536,870,912; 268,435,456
7,8,9, 22(*),23(*),24(*)	10	0%; 50%; 75%	1,073,741,824; 536,870,912; 268,435,456
10,11,12, 25(*),26(*),27(*)	30	0%; 50%; 75%	1,073,741,824; 536,870,912; 268,435,456

(*) Service managed by OTT

4.4 Results

Table 2 shows the results of simulations for the study cases. Results for class 2 are marked with (*).

5 Linear Models Fit

Multiple variable regression model is used for the identification of the relationship of QoE metrics Page Response Time (PRT), Video Response Time (VRT) with QoS KPIs, namely end-to-end Delay (DEL), Packet Loss Rate (PLR) and Throughput (THR). MNO-managed and OTT-managed classes are separately analyzed. In order to assess the regression model, QoS KPIs are considered as variable predictors (independent variables), QoE is the corresponding response (dependent variable). Since the regression model with one variable predictor (one QoS KPI) and one response (QoE) gives unsatisfying scatter plot of residuals a Multivariate Linear Regression (MLR) model has been here applied. A generic MLR model is represented by

$$QoE = \beta_0 + \beta_1 * DEL + \beta_2 * PLR + \beta_3 * THR + \varepsilon_i \quad (1)$$

where β_0, β_1, β_2,…, β_n are the unknown regression coefficients and $\varepsilon i \sim (N,\sigma 2)$ is the error of each observed response. QoE as the response (dependent variable) may be one PRT or VRT, while DEL, PLR, THR are the independent variables.

Table 2. Simulation Results.

Scenario no.	*Features Range [Min;Max]*				
	e2e Delay [s]	***PLR [%]***	***LTE THR [Mbps]***	***PRT***	***VRT***
1,2,3, 16(*),17(*), 18(*)	[0.08; 0.09]; [0.53; 0.91] (*)	[0.05; 0.91]; [0.67; 1.46] (*)	[3.20; 3.30]; [0.50; 0.80] (*)	[3.80; 4.18]; [7.85; 15.10] (*)	[5.28; 5.90]; [33.71; 43.30] (*)
4,5,6, 19(*),20(*), 21(*)	[0.08; 0.09]; [0.32; 0.58] (*)	[0.25; 0.43]; [6.14; 13.08] (*)	[2.60; 2.78]; [0.48; 0.63] (*)	[4.70; 4.73]; [9.09; 13.90] (*)	[8.73; 9.83]; [44.57; 49.77] (*)
7,8,9, 22(*),23(*), 24(*)	[0.08; 0.09]; [0.31; 0.61] (*)	[0.18; 0.57]; [12.19; 21,82] (*)	[1.88; 2.03] ; [0.65; 0.73] (*)	[6.13; 6.25]; [11,64; 15,71] (*)	[13.15; 13.25]; [39.14; 50.08] (*)
10,11,12, 25(*),26(*), 27(*)	[0.76; 0.77]; [2.52; 3.34] (*)	[0.27; 0.50]; [16,88; 22,99] (*)	[1.00; 1.05]; [0.43; 0.60] (*)	[9.21; 9.63]; [8.94; 17.09] (*)	[30.80; 31.40]; [40.20; 53.55] (*)
13,14,15, 28(*),29(*), 30(*)	[0.74; 0.82]; [4.33; 5.18] (*)	[11.34; 13.35]; [25,31; 38,56] (*)	[0.57; 0.59]; [0.53; 0.70] (*)	[10.66; 12.41]; [7.31; 11.52] (*)	[50.98; 52.95]; [41.03; 53.96] (*)
1,2,3, 16(*),17(*), 18(*)	[0.08; 0.09]; [0.53; 0.91] (*)	[0.05; 0.91]; [0.67; 1.46] (*)	[3.20; 3.30]; [0.50; 0.80] (*)	[3.80; 4.18]; [7.85; 15.10] (*)	[5.28; 5.90]; [33.71; 43.30] (*)
4,5,6, 19(*),20(*), 21(*)	[0.08; 0.09]; [0.32; 0.58] (*)	[0.25; 0.43]; [6.14; 13.08] (*)	[2.60; 2.78]; [0.48; 0.63] (*)	[4.70; 4.73]; [9.09; 13.90] (*)	[8.73; 9.83]; [44.57; 49.77] (*)
7,8,9, 22(*),23(*), 24(*)	[0.08; 0.09]; [0.31; 0.61] (*)	[0.18; 0.57]; [12.19; 21,82] (*)	[1.88; 2.03] ; [0.65; 0.73] (*)	[6.13; 6.25]; [11,64; 15,71] (*)	[13.15; 13.25]; [39.14; 50.08] (*)
10,11,12, 25(*),26(*), 27(*)	[0.76; 0.77]; [2.52; 3.34] (*)	[0.27; 0.50]; [16,88; 22,99] (*)	[1.00; 1.05]; [0.43; 0.60] (*)	[9.21; 9.63]; [8.94; 17.09] (*)	[30.80; 31.40]; [40.20; 53.55] (*)
13,14,15, 28(*),29(*), 30(*)	[0.74; 0.82]; [4.33; 5.18] (*)	[11.34; 13.35]; [25,31; 38,56] (*)	[0.57; 0.59]; [0.53; 0.70] (*)	[10.66; 12.41]; [7.31; 11.52] (*)	[50.98; 52.95]; [41.03; 53.96] (*)

(*) Service managed by OTT

5.1 MNO-managed class

5.1.1 PRT as the QoE Metric

The estimated regression model is as in

$$PRT = 9.357 + 2.368 * DEL + 9.619 * PLR - 1.749 * THR_i \quad (2)$$

This MLR model provides a low Root Mean Squared Error (RMSE) value equal to 0.443, an R-squared value equal to 0.982 and an Adjusted R-Squared value equal to 0.97: i.e. quite a good fit. The overall p-value is 6.15e-10. It allows us to reject the null hypothesis. Figure 4 shows relevant statistical indicators for this model: the fit looks actually good, while residuals are well distributed and adequately symmetric.

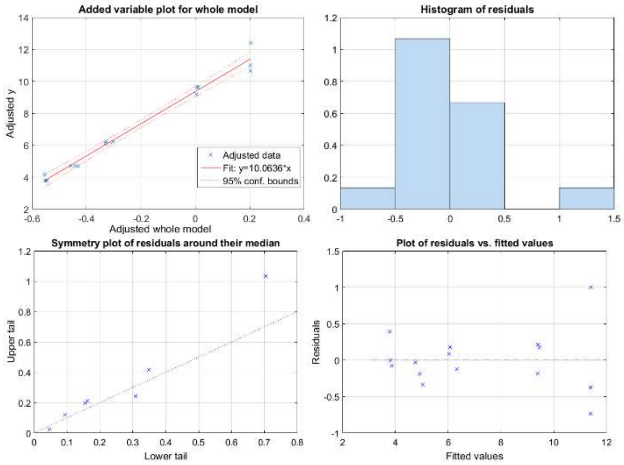


Figure 4. Statistical indicators for MLR of PRT in MNO-managed class. From upper leftmost to right: plot for whole model vs samples; histograms of residuals; symmetry plot of residuals around their median; residuals vs fitted values.

5.1.2 VRT as the QoE Metric

The estimated regression model is as in

$$VRT = 22.24 + 18.56 * DEL + 148.7 * PLR - 5.660 * THR \quad (3)$$

This MLR model provides a low Root Mean Squared Error (RMSE) value equal to 0.802, an R-squared value equal to 0.998 and an Adjusted R-squared equal to 0.986: i.e. quite a good fit. The overall p-value is 1.09e-15. It allow us to reject the null hypothesis. Figure 5 shows relevant statistical indicators for this model: the fit looks actually good, while residuals are well distributed and adequately symmetric.

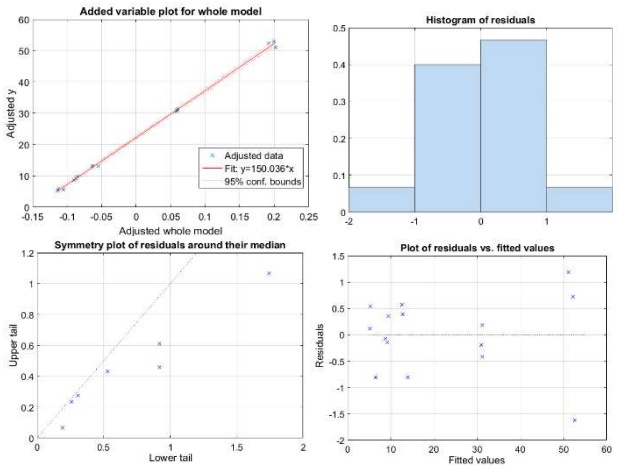


Figure 5. Statistical indicators for MLR of VRT in MNO-managed class. From upper leftmost to right: plot for whole model vs samples; histograms of residuals; symmetry plot of residuals around their median; residuals vs fitted values.

5.1.3 Cluster Analysis for MNO-managed class

Scenarios of the MNO-managed class, expressed in the 5 dimensional feature space, consisting of two QoE metrics and four QoS KPIs, have been clustered in three clusters, thanks to the well-known k-means technique [15]. Separation between clusters is fair as shown by the plot on the first two principal components (Figure 6).

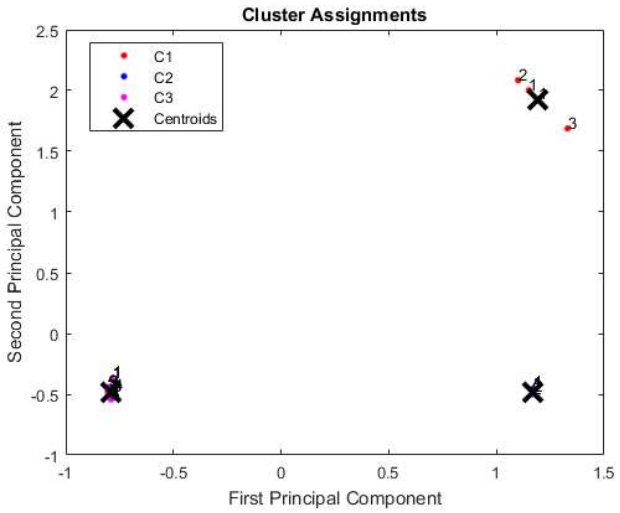


Figure 6. Clustering of scenarios for the MNO-managed class in the joint space of QoE metrics and QoS KPIs. Plot is on the plane of the first two principal components.

Meaning of the clusters is expressed in Table 3, which reports the position of the centroids of the three identified clusters in the 5-dim space. The quality of this clustering is evident from the changes of QoE between centroids: the VRT has the worst performance in C1, improves of 4.5 dB for C2, and 14.9 dB for C3. Same holds for PRT, though with lower gains (1.6 and 7.2 dB). As a conclusion, clustering well depicts differences between the scenarios in the MNO-managed class.

Table 3. Cluster Centroids generated by k-Means for MNO-managed Scenarios.

Cluster Centroid Id.	*DEL [s]*	*PLR [%]*	*THR [Mbps]*	*PRT [s]*	*VRT [s]*
1	0.7724	12.54 %	0.5783	11.36	52.05
2	0.7641	0.3600 %	1.025	9.489	31.01
3	0.0833	0.3700 %	2.650	4.944	9.339

5.2 OTT-managed class

5.2.1 PRT as the QoE Metric

The estimated regression model is as in

$$PRT = 7.347 - 0.813 * DEL + 13.00 * PLR + 6.796 * THR \quad (4)$$

This MLR model is affected by an excess Root RMSE at 3.11 and an insufficient R-squared value at 0.13, while Adjusted R-Squared is meaningless (-0.08). The overall p-value is 0.591. Thus we don't reject the null hypothesis. Let also remark that also the positive gain PRT/THR makes no sense. Figure 7 shows some evidences of the fact that the MLR model poorly describes the potential relationship QoE vs QoS in case of PRT.

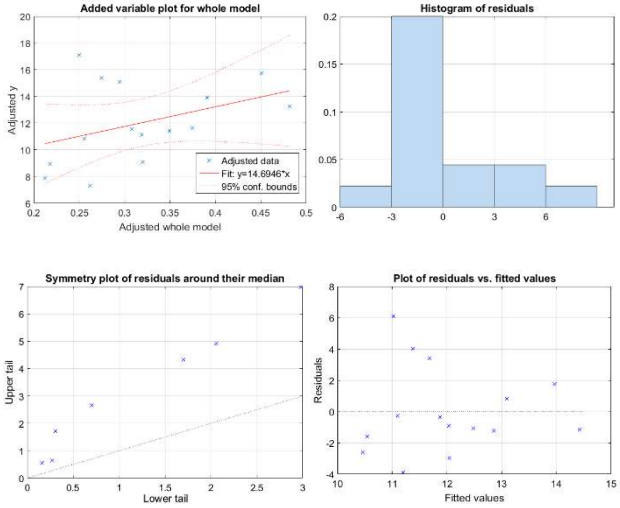


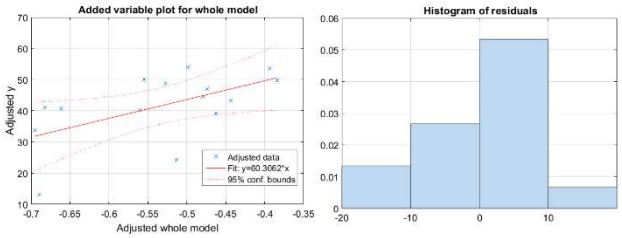
Figure 7. Statistical indicators for MLR of PRT in OTT-managed class. From upper leftmost to right: plot for whole model vs samples; histograms of residuals; symmetry plot of residuals around their median; residuals vs fitted values.

5.2.2 VRT as the QoE Metric

The estimated regression model is as in

$$VRT = 73.78 - 3.200 * DEL + 32.46 * PLR - 50.72 * THR \quad (5)$$

However, the MLR model is even worse than (4) (RMSE=10.2 and Adjusted R-Squared=0.14). The overall p-value of 0.211 makes the null hypothesis really likely. Figure 8 shows some evidences of the fact that the MLR model poorly describes the potential relationship QoE vs QoS in case of VRT.



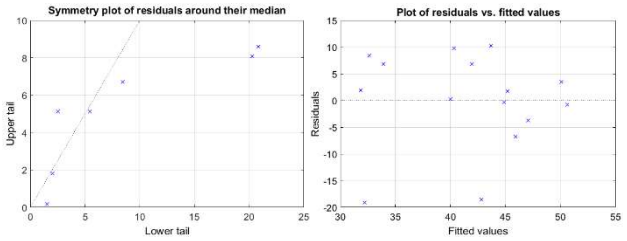


Figure 8. Statistical indicators for MLR of VRT in OTT-managed class. From upperleftmost to right: plot for whole model vs samples; histograms of residuals; symmetry plot of residuals around their median; residuals vs fitted values.

5.2.3 Cluster Analysis for OTT-managed class

The statistical indicators in the OTT-managed class do not allow to take the fit as trusted. It means that a linear relationship between QoE KPIs and QoS metrics cannot really be ascertained. Cluster analysis can gives some more insight on relationships between scenarios of the OTT-Managed class, represented in the 5 dimensional feature space, consisting of two QoE metrics and three QoS KPIs. Figure 9 plots two clusters identified thanks to the k-means algorithm [15] on the plane identified by the first two principal components.

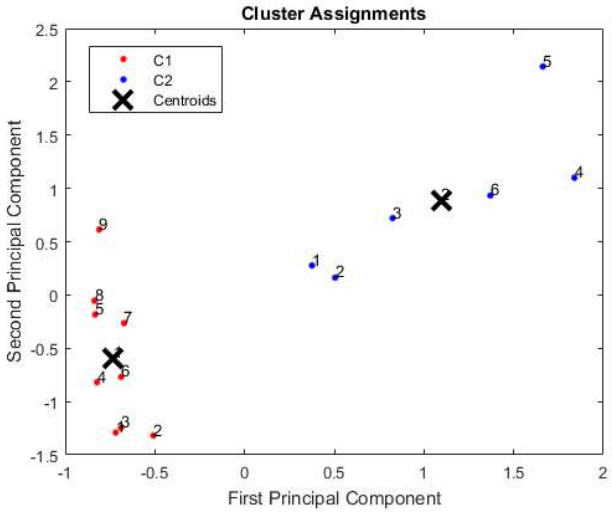


Figure 9. Identified clusters for the OTT-managed class including jointly both PRT and VRT as QoE metrics.

Separation between the two clusters is fair as per Figure 9. Though meaning of this separation is quite different as for the MNO-managed class. Table 4 reports the position of the centroids of the three identified clusters in the 5-dim space. It is straightforward noting that the two cluster centers are different as to the QoS KPIs, while they are nearly undistinguishable as for the QoE KPIs. Therefore, clustering reveals under a different perspective that QoE KPIs are not predictable starting from the QoS metrics.

Table 4. Cluster Centroids generated by k-Means for MNO-managed Scenarios.

Cluster Centroid Id.	*DEL [s]*	*PLR [%]*	*THR [Mbps]*	*PRT [s]*	*VRT [s]*
1	0.5026	8.610	0.6473	12.09	41.58
2	3.829	24.84	0.5720	11.89	41.45

6 Discussion and Final Remarks

The procedure of QoE metrics estimation out of QoS KPIs measured at network level is a consolidated process, and still a due attempt, regardless the complexity of networks and service architectures. Here we attempted for LTE networks the estimation of QOE vs QoS mapping thanks to MLR technique, with two distinct classes of scenarios. Our approach worked well as long as the MNO has been able to perform an end-to-end management of the service delivered to the end user (MNO-managed class). It is possible to give a reliable estimation of QoE out of the QoS due to the MNO. This case may represent an incentive for OTT, who is supposed to have the ownership of contents, to cooperate with the MNO within the service delivery process. In the OTT-managed class the MNO loses the end-to-end control and the MLR technique turns out to give unsatisfactory results. In our understanding this is due to the fact that there is no-one with a full end-to-end management of the HTTP service.

Our results bring us to two key considerations. First, regression models are suitable for QoE vs QoS mapping if the QoS KPIs measured by MNO are the only factors that determine QoE. Second, if other factors out of the MNO's control affects the service delivery process (as the IP cloud considered in this work), the identification of the QoE vs QoS mapping comes to an unsatisfactory end.

A future step is to determine for the OTT-Managed class the minimal Service Level Agreement (SLA) of IP cloud that assures the existence of a QoE vs QoS (on the MNO side) mapping. This enables the OTT to negotiate with the MNO consistent QoS KPIs for the intended QoS metrics. Next steps regard how to assure that reliable mappings can be utilized on both sides, MNO and OTT, to deliver a quality-safe service to the end-user, within a sustainable business model. Impact of the QoE vs QoS mapping on costs and economic measures has to take part in the analysis

References

3GPP TS 23.401- 2007 GPRS Enhancements for E-UTRAN Access. Available via www.3gpp.org [accessed March 29, 2018].

3GPP TS 23.402-2009 Architecture enhancements for non-3GPP Accesses. Available via www.3gpp.org [accessed March 29, 2018].

M. Alreshoodi and J. Woods. Survey on QoE/QoS correlation models for multimedia services. *International Journal of Distributed and Parallel Systems*, 3: 53-72, 2013.

M. Alreshoodi and J. Woods. An empirical study based on a Fuzzy Logic System to Assess the QoS/QoE Corelation for Layered Video Streaming. In Proceedings *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications*, pages 180-184, 2013.

Ericsson Mobility Report. Ericsson Ltd, 2015.

ITU-T Recommendation, ITU-T Rec. P.800.1: Mean Opinion Score Terminology.

M. Fiedler and T. Hossfeld, and P. Tran-Gia. A Generic Quantitative Relationship between Quality of Experience and Quality of Service. *IEEE Network*, 2, pages 36-41, 2010.

S. Khorsandroo and R. Md Noor, and S. Khorsandroo. Stimulus-Centric versus Perception-Centric Relations in Quality of Experience Assessment. In Proceedings *IEEE Wireless Telecommunications Symposium*, London, pages 1-6, 2017.

H. G. Msakni and H. Youssef. Is QoE estimation based on QoS parameters sufficient for video quality assessment?. In Proceedings *9th International Wireless Communications and Mobile Computing Conference*, pages 538-544, 2013.

M. Siller and J. Woods. QoS arbitration for improving the QoE in multimedia transmission. In Proceedings *IEEE International Conference on Visual Information Engineering,* pages 238-241, 2003.

T.H Truong and Tai-Hung Nguyen, and Huu-Thanh Nguyen. On relationship between Quality of Experience and Quality of Service metrics for IMS-based IPTV networks. In Proceedings *IEEE International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future*, pages 1-6, 2012.

J.Urrea and N. Gaviria. Statistical performance evaluation of P2P video streaming on multi-hop wireless networks. In Proceedings *20th Symposium on Signal Processing, Images and Computer Vision*, pages 1-6, 2015.

A. Vizzarri and S. Forconi. Review of Studies on End-to-End QoS in LTE Networks. In Proceedings *AEIT Congress*, pages 1-6, 2013.

A. Vizzarri. Analysis of VoLTE end-to-end quality of service using OPNET. In Proceedings *European Symposium on Computer Modeling and Simulation*, pages 452 – 457, 2014.

A. Vizzarri. Analysis of VoIP Over LTE End-To-End Performances in Congested Scenarios. In Proceedings *Artificial, Intelligence, Modelling and Simulation*, pages 393-343, 2014.

Wang, B. and Wen, X. and Yong, S., and Wei. A New Approach Measuring Users' QoE in the IPTV. In Proceedings *IEEE Pacific-Asia Conference on Circuits, Communications and Systems*, pages 453-456, 2009.

Simulation of VoLTE Services for QoE Estimation

Alessandro Vizzarri[1] Fabrizio Davide[2]

[1]Department of Enterprise Engineering, University of Rome Tor Vergata, Italy,
alessandro.vizzarri@uniroma2.it

[2]Department of Innovation and Information Engineering, Guglielmo Marconi University, Italy,
f.davide@unimarconi.it

Abstract

One the most important features of a Long Term Evolution (LTE) system is the high transmission data rate in downlink and in uplink. This is not sufficient for a good Quality of Experience (QoE) perceived by the end user. The Mobile Network Operator (MNO) has to adopt appropriate techniques for an effective management of the Quality of Service (QoS) not only for bandwidth-consuming applications as video streaming but also for voice application as Voice Over LTE (VoLTE). These techniques can be based on the QoE/QoS correlation especially in case of a delaysensitive application as VoLTE. This paper formulates a method for the QoE estimation starting from QoS informations available at network level. Different scenarios are simulated using OPNET software tool. Results are statistically analyzed using regression cluster analysis techniques. Mathematical functions representing relationship between QoE/QoS metrics are identified.

Keywords: LTE, QoS, QoE, VoLTE, key performance indicators, regression, cluster

1 Introduction

Wireless telecommunication networks have enabled broadband applications with very high throughput. With the introduction of Long Term Evolution (LTE) the download data rate can reach 100 Mbps, while the upload data rate may be up to 50 Mbps (3GPP,2008). The Quality of Service (QoS) have been so strongly impacted by these new important capabilities that an efficient management of the LTE network is needed in order to guarantee acceptable levels of Quality of Experience (QoE) to the end users. Together with data applications, the voice application called Voice Over LTE (VoLTE) is also to be carefully managed. This implies the importance for the Mobile Network Operator (MNO) to make a reliable QoE estimation. MNOs need to integrate LTE native QoS features with other techniques that consider the entire communication chain. One approach is to correlate the QoS measured at network level to the QoE perceived by end user. In the scientific literature several mathematical models have been defined.

This work is focused on the QoE estimation for VoLTE application on the basis of network QoS indicators, called Key Performance Indicators (KPIs). The paper presents in Sect. II an overview of the related work and the proposed approach for QoE/QOS correlation in Sect. III. In Sect. IV a simulation activity is detailed and the results are analyzed in the Sect. V. Sect. VI resumes the main conclusions.

2 Related Works

Reference (3GPP, 2008) introduces the end-to-end QoS reference architecture for LTE systems as standardized by ETSI together with the basic management functions. A. Vizzarri et alii in (Vizzarri, 2013) present a review of most important papers on end-to-end QoS approach in LTE networks. Horvath et alii in (Horvath, 2013) present an innovative signalling protocol named LQSIG for the resource reservation.

A first attempt for correlating quality informations of the LTE network to those of the application is made by S. Shen et alii in (Shen, 2011). They propose a performance framework based on the mapping of the Class of Service (CoS) to the QoS. Margoc et alii in (Margoc, 2013) analyze QoS in LTE systems in order to analyze the better performances for higher priority services. Policies and strategies for priority service allocation are left to operators. This is also confirmed by Medbo et alii in (Medbo,2009), where different type of data traffic over LTE network are analyzed, e.g. VoIP and HTTP web browsing/video streaming.

In (Alreshoodi,2013) QoS/QoE correlation is studied assuming QoS as a source of disturb for QoE. In (Alreshoodi,2013) fuzzy-logic approach is proposed for QoS/QoE mapping. In (Truong, 2012) the QoS/QoE mapping for IPTV service is made on the basis of the Mean Opinion Score (MOS) (intended as the QoE metric) delay, jitter and Packet Loss Rate (intended as the QoS metrics). A. Vizzarri in (Vizzarri, 2014) analyzes the impact of the voice codec on the end-to-end QoS for a VoLTE service. In [Vizzarri, 2014) A. Vizzarri studies the impact of the network congestion (in terms of link utilization) on end-to-end QoS for a VoLTE service.

3 Proposed Approach

3.1 Methodology

LTE system enables fully IP-based applications thanks to data transmitted through Packet Switching (PS) paths. That implies VoLTE to be treated as an IP-based application (3GPP, 2008). IP Multimedia Sub System (IMS) (3GPP, 2008) and Session Initiation Protocol (SIP) are integrated with LTE network nodes (IETF, 2005).

Figure 1 shows the LTE logical architecture supporting VoLTE service includes VoLTE User Equipment (UE), Evolved Universal Terrestrial Radio Access Network (E-UTRAN), Evolved Packet Core (EPC) and IMS Core Network.

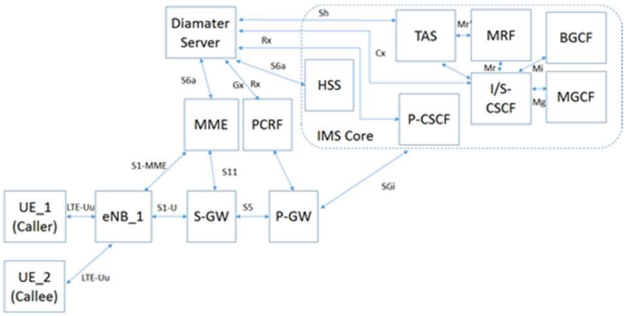


Figure 1. Logical architecture for VoLTE service [3GPP TS 23.002].

As a fully IP-based application, VoLTE is delivered over the LTE in a best effort modality. That implies the Mobile Network Operator (MNO) has to manage in order to guarantee acceptable levels of both QoS measured at the network level and QoE perceived by the end user at the application level. QoE is usually represented by a subjective measure called Mean Opinion Score (MOS). MOS is a scalar variable that measures the degree of service acceptance by the end user on range from 1 (worst case) to 5 (best case) (ITU, 2016). A MOS value equal to 5 is indicative of an Excellent Quality (imperceptible impairment), 4 of a Good Quality (perceptible but non annoying impairment), 3 of a Fair Quality (slightly annoying impairment), 2 of a Poor Quality (annoying impairment) and 1 of a Bad Quality (very annoying impairment). MOS value is derived from to R factor provided by ITU E-Model (ITU,2008) which takes in account several factors impacting the QoS, e.g. choose of voice codec, transmission delay, etc.

KPIs measured at the network level are typically Delay (DEL), Jitter (JIT) and Packet Loss Rate (PLR) (Yu et alii, 2007). The delay is represented as the amount of time a packet sent by source (caller) takes to reach destination (callee). Jitter is the variation in the time between packets arriving, caused by network congestion or route changes. Negative effects of Delay and Jitter are represented by phenomena of voice echoes. A high value of PLR can produce overlapping of words with a strong negative impact on voice intelligibility.

The present paper is mainly focused on the study and the identification of mathematical models for QoS/QoE correlation in different VoLTE realistic scenarios. Considered metrics are:

- QoE metrics: MOS;
- QoS KPIs: end-to-end Delay (e2eDEL), Jitter (JIT) and Packet Loss Rate (PLR).

Correlation techniques here adopted by the authors are statistical regression and cluster analysis.

3.2 Network models

In order to build realistic scenarios for VoLTE, network impairments are included. They are modelled on the basis of two constraints:

- Presence of mixed traffic: VoLTE application is delivered over LTE network together with HTTP web browsing application;
- Presence errors on end-to-end transmission: they are modelled inserting an IP cloud which introduces both additive delay (0.1 seconds) and IP packet discard ratio (1%).

On the basis of the VoLTE logical architecture shown in Figure 1, Figure 2 reports four different LTE network models, built in order to model realistic conditions for the VoLTE application and compute QoS and QoE metrics.

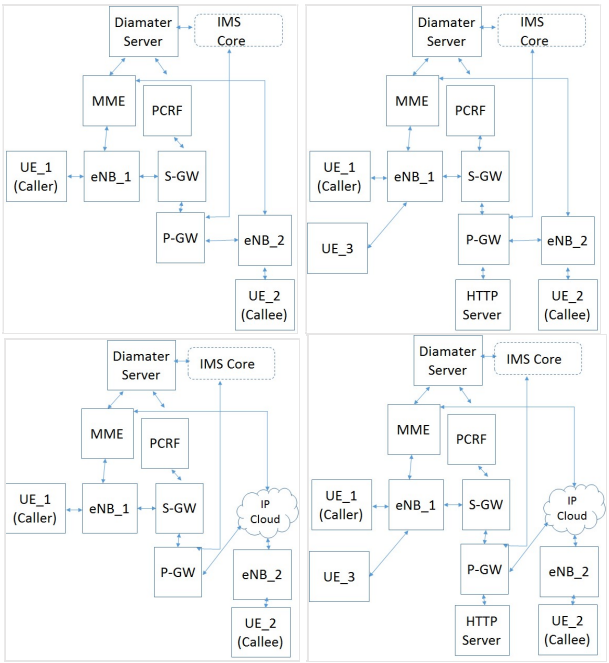


Figure 2. Four LTE network models identified to model realistic conditions for for VoLTE application. From the upper leftmost to right: UE for VoLTE without IP cloud; two UEs for VoLTE and HTTP browsing without IP cloud; One UE for VoLTE with IP cloud; two UEs for VoLTE and HTTP browsing with IP cloud.

The first model is related to a classic VoLTE call between two LTE UEs: a caller (UE_1) and a callee (UE_2). It is the situation with a single LTE application (VoLTE) without any network impairment. The second model is related to a situation with two different services delivered over the same LTE network. VoLTE is the first application between UE_1 and UE_2; HTTP web browsing is the second one and it is performed by UE_3 contacting an HTTP server. UE_1 (VoLTE caller) and UE_3 are attached to the same eNodeB (eNB_1). The third model is an extension of the first one. VoLTE application is affected by only the insertion of an IP cloud across the end-to-end transmission chain. The fourth model is an extension of the second one. Here VoLTE application is affected by both the presence of IP cloud and the presence of a second application (HTTP web browsing) performed by another user.

4 Simulation

All scenarios are simulated using the LTE network model provided by OPNET 17.5 PL6.

4.1 OPNET Settings

Both the antenna gain equal to -1 dBi and the receiver sensitivity to -200 dBm characterize the UE. eNodeB transmission mode is FDD Duplex Mode. Link type among LTE network nodes is PPP D3: the data rate is 44.736 Mbps. Voice codec is GSM EFR with one voice frame per packet. EPS bearer has a QoS Class Identifier (QCI) equal to 1 (GBR) and an Allocation and Retention Priority (ARP) equal to 1. Through the HTTP application, UE_3 can download 1 KB web page, n. 5 medium images with dimension up to 2 KB and two short videos with dimension up to 350 KB. The simulation period is equal to 3 minutes for all scenarios. The simulation area is a typical campus area (100 Km^2).

4.2 Assumptions on users and networks

According to the four network models defined in the Sect. III, we identified 48 scenarios to be simulated.

The scenario set includes two tunable parameters:

- S1 link capacity: starting form 100% (corresponding to 1,073,741,824 bps), the link capacity is decreased to 75%, 50% and 30%;
- eNodeB Bandwidth: 5 MHz, 10 MHZ, 20 MHz.

Table 1 reports scenarios grouped in four subsets, and their parameters.

4.3 Results

Table 2 gives a view of simulation results per homogeneous groups of scenarios, in terms of related QoS metrics and QoE metric. As a first remark, MOS is acceptable for the first two subsets (fair quality), and too low for the third (poor quality) and fourth (bad quality) subsets.

Table 1. Scenario Configurations.

Subset No.	*Scenario No.*	*LTE Service Type*	*S1 Link Capacity [%]*	*eNB Band [MHz]*	*Impairments due to IP Cloud*	
					Packet Discard Ratio [%]	*Packet Latency [s]*
1	1-12	VoLTE	100; 75; 50; 30	5; 10; 20	Not present	Not present
2	13-24	VoLTE + HTTP Browsing	100; 75; 50; 30	5; 10; 20	Not present	Not present
3	25-36	VoLTE + IP Cloud	100; 75; 50; 30	5; 10; 20	1	0.1
4	37-48	VoLTE + HTTP Browsing + IP Cloud	100; 75; 50; 30	5; 10; 20	1	0.1

Table 2. Simulation Results.

Subset No.	*Scenario No.*	*QoS Metrics*			*QoE Metrics*
		Delay [s]	*Jitter [s]*	*PLR [%]*	*MOS*
1	1-12	[0.11; 0.12]	[0.06; 0.11]	[0.21; 0.53]	[3.46; 3.70]
2	13-24	[0.12; 0.13]	[0.09; 0.13]	[1.16; 1.37]	[2.90; 3.50]
3	25-36	[0.19; 0.21]	[0.15; 0.19]	[7.40; 8.85]	[2.26; 2.50]
4	37-48	[0.26; 0.27]	[0.21; 0.27]	[21.60; 26.06]	[1.45; 1.70]

5 Statistical Analysis

5.1 Single Variable Non Linear Regression Model

A generic regression model is representable as

$$Y = f(x) + \varepsilon_i \quad (1)$$

where Y is the response (dependent variable) to the predictor x (independent variable), with $\varepsilon_i \sim (N,\sigma_i^2)$ the error for each observation. The relationship function f(x) can be linear or not. Based on available dataset the model of choice for regression is non-linear, and

specifically exponential (Alreshoodi, 2013). In case of MOS as response and a KPI as predictor, (1) becomes

$$QoE = a * exp(-b * QoS) + c \qquad (2)$$

where a and b are numerical coefficients. QoS may be Delay, Jitter and PLR in turn.

5.1.1 MOS vs DEL fitting.

In case of fitting between MOS and Delay, the regression model is

$$MOS = -13.96 * exp(-1.133 * DEL) - 8.783 \qquad (3)$$

Figure 3 gives a graphic representation of (3). The fitting quality is proved by a good value of R-squared (0.9487) (the adjusted R-Squared is quite near, i.e. 0.9464). Further the value of Root Mean Square Error (RMSE) is low enough, i.e. 0.185.

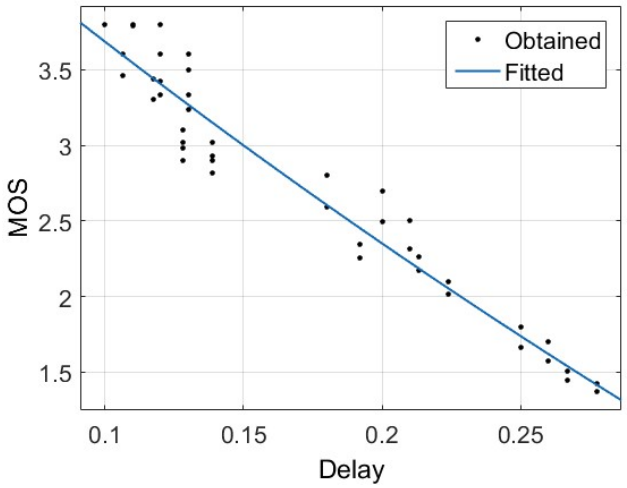


Figure 3. MOS to Delay regression model.

5.1.2 MOS vs JIT fitting.

In case Jitter is considered as a QoS, the regression model is estimated as

$$MOS = -94.21 * exp(-0.1219 * JIT) - 98.78 \qquad (4)$$

Figure 4 shows the plot for (4). R-squared value is as for (3), i.e. 0.9374 (as much as the adjusted R-squared, i.e. 0.9352). The RMSE value is 0.2034. We can conclude the fit to be adequate.

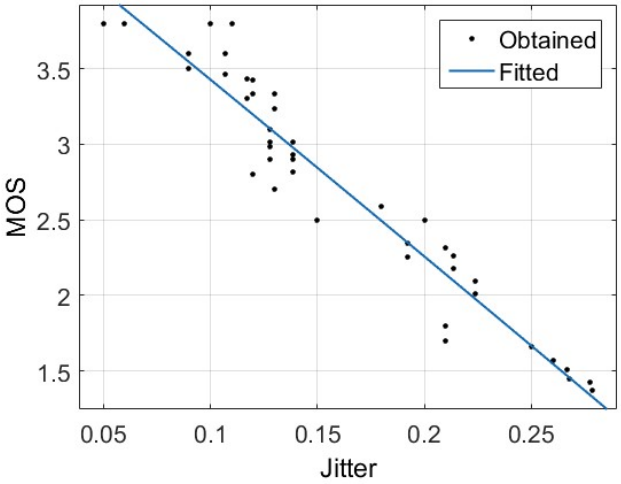


Figure 4. MOS to Jitter regression model.

5.1.3 MOS vs PLR fitting.

As far as PLR is considered, the regression model comes out to be

$$MOS = 2.198 * exp(-11.06 * PLR) + 1.391 \qquad (5)$$

Figure 5 shows graphically the regression model. Data exhibit a strong non linear character. R-square value is 0.8969, while RMSE value is 0.226. The quality of fitting is slightly lower than for (3) and (4).

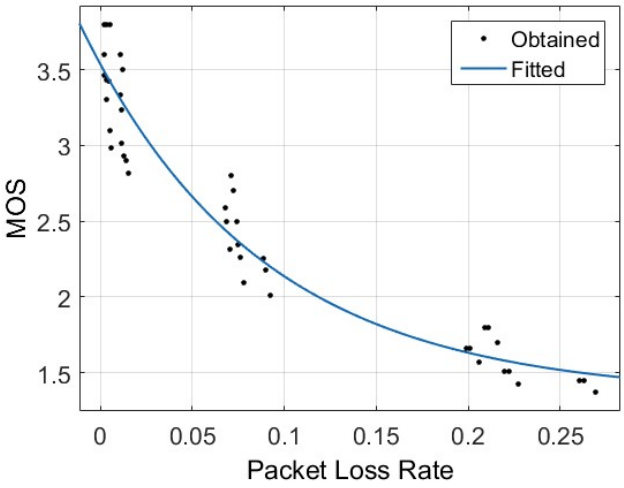


Figure 5. MOS to PLR regression model.

Further Figure 5 reports the existence of 3 clusters in the scenarios set. The clusters are linearly separable. This effect is similar in the (MOS, DEL) plane, and not evident in the (MOS, JIT) plane.

5.2 Multiple Variable Linear Regression Model

The previous fitting models are able to identify a (sometimes-strong) non-linear relationship between MOS and QoS KPIs taken once a time. These models give for VoLTE an idea of the QoE/QoS mapping but have a limited validity. In this section we look for a relationship between all QoS KPIs and QoE, a multi-variable approach is helpful. The Multiple variable regression (MLR) model is usually represented as

$$Y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_n * x_n + \varepsilon_i \qquad (6)$$

where β_0, β_1, β_2,…, β_n are the unknown regression coefficients and $\varepsilon_i \sim (N,\sigma_i^2)$ is the error for each observation.

As far as MOS is the response (dependent variable) and DEL, PLR, THR are the independent variables, from (6) we have

$$MOS = \beta_0 + \beta_1 * DEL + \beta_2 * JIT + \beta_3 * PLR \qquad (7)$$

For our dataset, the regression model comes out as

$$MOS = 4.824 - 6.958 * DEL - 5.533 * JIT - 0.1477 * PLR \qquad (8)$$

Figure 6 shows relevant statistical indicators for model (8): the fit looks actually good, and residuals are well distributed and adequately symmetric. The goodness of the regression model (8) is proved by R-squared quite high 0.973 (with an adjusted R-squared equal to 0.971) and a quite low value of RMSE, i.e. 0.136. The overall pValue is very low: 1.75e-34. This means the null hypothesis can be firmly rejected.

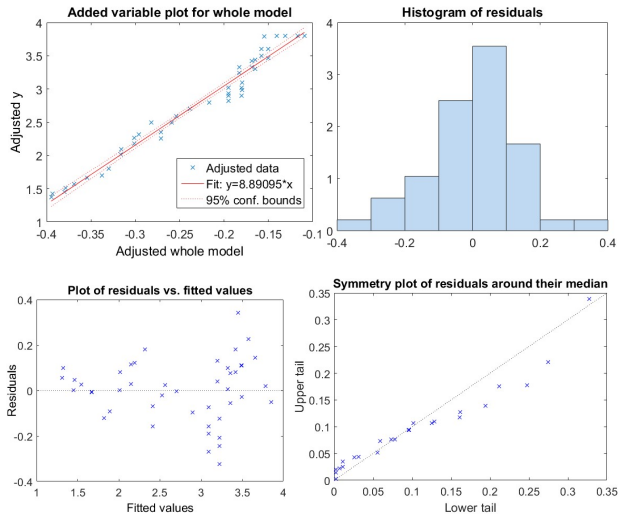


Figure 6. Statistical indicators for MLR in case of MOS, DEL, JIT and PLR. From upperleftmost to right: plot for whole model vs samples; histograms of residuals; symmetry plot of residuals around their median; residuals vs fitted values.

Table 3 presents the main statistical parameters for the coefficients in (7) and (8).

Table 3. MLR Results.

Coefficients	*Estimated Value*	*Standard Error (SE)*	*tStat*	*pValue*
β0 (intercept)	4.824	0.1339	36.035	2.67e-34
β_1	-6.958	1.36	-5.116	6.568e-06
β_2	-5.533	0.8751	-6.323	1.131e-07
β_3	-0.1477	0.6879	-0.2147	0.8310

5.3 Cluster Analysis

Besides the results coming out of the regression analysis, we need to gain a better insight in the QoE/QoS mapping. Clustering may provide that. Let us map the scenarios in the 4 dimensional feature space, whose dimensions are MOS, DEL, JIT and PLR. If we apply the well-known k-means technique [20] [21], we get three clusters as a result after the algorithm stabilization. Figure 7 shows the clusters projected onto the first two principal components. Clusters are linearly separable and fairly separated.

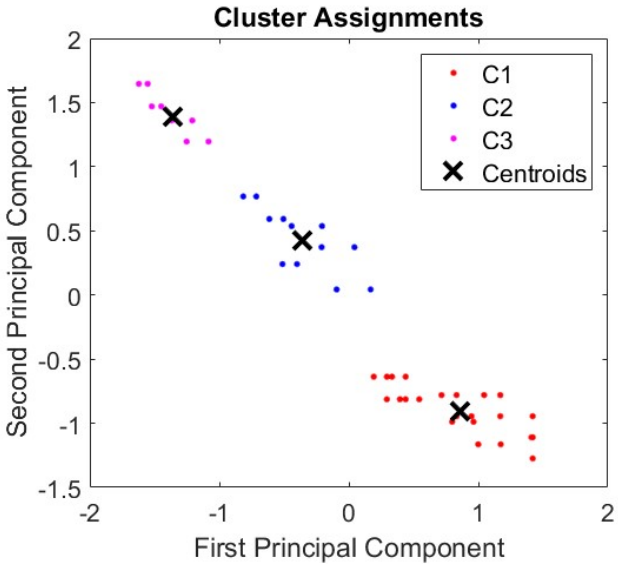


Figure 7. Clusters of VoLTE scenarios. Plot is a projection on the plane of the first two principal components. Centroids are marked as crosses.

Table 4 reports the position of the centroids of the three identified clusters in the 4-dim space. The utility of this clustering is evident from the changes of QoE between the centroids: MOS assumes the best value in C1 that consists of 24 scenarios, characterized by a single application: (VoLTE in the scenarios 1-12) or by

a mixed traffic, (VoLTE and HTTP web browsing applications, scenarios 13-24).

The MOS gets worse of -3.1 dB for C2 when the VoLTE application co-exists with end-to-end transmission errors due to an IP cloud (scenarios 25-36). The MOS decreases of -6.8 dB in C3 where both a mixed traffic (VoLTE and HTTP web browsing) and end-to-end transmission errors are present.

Table 4. MLR Results.

Cluster No.	*Scenario No.*	*MOS*	*Delay [s]*	*Jitter [s]*	*PLR [%]*
1	[1-24]	3.40	0.12	0.11	0.77
2	[25-36]	2.39	0.20	0.19	8.20
3	[37-48]	1.56	0.26	0.25	22.54

To be remarked that cluster C1 gathers scenarios from subsets 1 and 2. C2 is overlapping with subset 2 of the scenarios. C3 is overlapping with subset 3. Based on Table II, Table IV shows both this relationship and related cluster characteristics.

6 Conclusions

VoLTE is an important service for LTE networks since it is the basic voice service in the 4G wireless standard. Since LTE is a fully IP-based system, VoLTE application is delivered in the best effort modality. Mapping QoE/QoS is of key importance for the MNOs. Here we studied how the correlation between QoE and QoS can be expressed through mathematical models.

We designed a certain number of scenarios, that include as network impairments both mixed traffic (co-presence of VoLTE and HTTP web browsing), and end-to-end transmission errors (presence of an IP cloud). We analyzed results from simulation through different techniques: single variable regression, Multi Linear Regression (MLR) and cluster analysis.

From the single variable nonlinear regression, we learnt that DEL and JIT have linear models. On the contrary, the PLR model exhibits a stronger non-linearity, which is the reason why presence of clusters in data has been anticipated from Figure 5.

The multivariable linear regression came to results comparable with the single variable regression for dependence from DEL and JIT. This is straightforward from the linearization of (3) and (4). Further, model (8) expresses the joint influence of all QoS metrics on MOS, with a good statistical quality. Nonetheless, the coefficient related to PLR has an unsatisfying pValue. This may be motivated by the irrelevance of the variable (that is the likelihood of the null hypothesis). Further, it may be a signature of statistical instability when a nonlinear behavior is forced to accommodate into a linear model. Related work demonstrated PLR to be a relevant predictor for MOS. Thus we consider (8) relevant under a theoretical viewpoint, but of limited use as far as a large PLR variation is considered.

The cluster analysis showed that the designed scenarios naturally group into three distinct clusters. Each of them has different features in terms of QoE and QoS metrics. As a first result, the authors observed that scenarios of C3 have an unacceptable MOS, due to errors of end-to-end transmission (modeled through IP cloud). About the role of PLR, clustering reveals a strong capability of PLR to predict existence of cluster (compare Figure 3 and Figure 7), and a very high discrimination between the clusters based on the PLR value (see Table IV)

Future works will be focused on the mapping QoE/QoS for LTE networks with different scenarios and design parameters. Applications in the scope will be video streaming and HTTP web browsing.

References

M. Alreshoodi and J. Woods. Survey on QoE/QoS correlation models for multimedia services. *International Journal of Distributed and Parallel Systems*, 3: 53-72, 2013.

M. Alreshoodi and J. Woods. An empirical study based on a Fuzzy Logic System to Assess the QoS/QoE Correlation for Layered Video Streaming. In Proceedings *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications*, pages. 180-184, 2013.

G. Horvath. End-to-end QoS Management Across LTE Networks. In *Proceedings 21st International Conference on Software, Telecommunications and Computer Networks*, pages 1-6, 2013.

IETF RFC 4028. Session Timers in the Session Initiation Protocol. Available via https://www.ietf.org/ [accessed March 29, 2018].

ITU Recommendation P.800.2 Series P: Terminals and Subjective and objective assessment methods. Available via https://www.itu.int/ [accessed March 29, 2018].

ITU Recommendation G.109 Definition of categories of speech transmission quality. Available via https://www.itu.int/ [accessed March 29, 2018].

A. Margoc. Quality of Service in Mobile Networks. In Proceedings *55th International Symposium ELMAR*, pages 263-267, 2013.

J. Medbo. Propagation channel impact on LTE positioning accuracy: A study based on real measurements of observed time difference of arrival. In Proceedings *20th International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 2213-2217, 2009.

J. A. Rice. *Mathematical Statistics and Data Analysis*. 3rd Edition, Belmont, CA, Thomson Brooks/Cole, 2009.

S.Shen. End-to-End QoS performance management across LTE networks. In Proceedings *3th Asia-Pacific Network Operations and Management Symposium*, pages 1-4, 2009.

Release 8 V0.0.3, Overview of 3GPP Release 8: Summary of all Release 8 Features. 2008. Available via www.3gpp.org/ [accessed March 29, 2018].

3GPP TS 123.207 Digital cellular telecommunications system; Universal Mobile Telecommunications System; LTE; End-to-end Quality of Service concept and architecture. Available via www.3gpp.org/ [accessed March 29, 2018].

3GPP TS 23.002 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects; Network architecture. Available via www.3gpp.org/ [accessed March 29, 2018].

3GPP TS 29.228 Digital cellular telecommunications system; Universal Mobile Telecommunications System; LTE; IP Multimedia Subsystem Cx and Dx Interfaces; Signaling flows and message contents. Available via www.3gpp.org/ [accessed March 29, 2018].

3GPP TS 32.260 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Telecommunication management; Charging management; IP Multimedia Subsystem charging. Available via www.3gpp.org/ [accessed March 29, 2018].

T.H Truong and Tai-Hung Nguyen, and Huu-Thanh Nguyen. On relationship between Quality of Experience and Quality of Service metrics for IMS-based IPTV networks. In Proceedings *IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future*, pages 1-6, 2012.

J.Urrea and N. Gaviria. Statistical performance evaluation of P2P video streaming on multi-hop wireless networks. In Proceedings *20th Symposium on Signal Processing, Images and Computer Vision*, pages 1-6, 2015.

A. Vizzarri and S. Forconi. Review of Studies on End-to-End QoS in LTE Networks. In Proceedings *AEIT Congress*, pages 1-6, 2013.

A. Vizzarri. Analysis of VoLTE end-to-end quality of service using OPNET. In Proceedings *European Symposium on Computer Modeling and Simulation*, pages 452 – 457, 2014.

A. Vizzarri. Analysis of VoIP Over LTE End-To-End Performances in Congested Scenarios. In *Proceedings Artificial, Intelligence, Modelling and Simulation*, pages 393-343, 2014.

J. Yu and I. Al-Ajarmeh. Call admission control and traffic engineering of VoIP. In Proceedings *2nd International Conference on Digital Telecommunications*, pages 11–16, 2007.

Constructive Assessment Method for Simulator Training

Laura Marcano Tiina Komulainen
Department of Electronic Engineering, Oslo and Akershus University College of Applied Sciences (HiOA), Norway
{Laura.Marcano,Tiina.Komulainen}@hioa.no

Abstract

Industrial operator assessment is a very controversial subject in the scientific community, as determining the most suitable, objective and effective means of giving feedback on an operator's performance is a great challenge. This paper presents a proposal on assessment methods for simulation training. The development is based on the results from simulator training courses held at Oslo and Akershus University College of Applied Sciences (HiOA) from 2010 to 2014. The results and course evaluation were analyzed to identify where new methods could be applied that would lead to improvement. The method proposed consists of an automatic assessment procedure, which will give feedback to the simulator course participants during the simulator session and help the students to achieve the learning outcomes. The proposed method will be tested in the simulator training courses at HiOA in spring 2017 and the results will be presented in a later paper.

Keywords: assessment, performance, operator, feedback, students, learning outcome

1 Introduction

1.1 Simulator training and performance assessment challenges

The evaluation of operators' performance represents a significant challenge for the process industry, as the appropriate assessment of operators' performance is of great importance to ensuring the right competencies and safe plant operations.

A recent study in the Norwegian oil and gas industry (Komulainen and Sannerud, 2014) reveals that only 30% of the respondents take exams after the simulation courses. The evaluation of the simulator trainee performance is based on the instructor's verbal feedback during the scenario and the instructor's verbal assessment after the scenario.

The automatic assessment tools available require the implementation of a specific sequence of actions for each scenario. The main criticism of automatic assessment is the high implementation and maintenance workload of the scenarios, the difficulty of implementing just one optimal sequence for complex scenarios, i.e. there can be many good alternative solutions, and the interpretation of operators' learning outcomes, competencies and skills from the figures generated by the automatic assessment system. Thus, the use of automatic assessment tools is not widespread in the Norwegian oil and gas industry.

Virtual laboratories i.e. complex process simulators, are important learning tools in modern engineering education; they are relevant to industrial practice, they facilitate collaborative, active learning among the students, and they are time and cost effective (Coble et al., 2010; Corter et al., 2011; Edgar et al., 2006; Komulainen and Løvmo, 2014; Martin-Villalba et al., 2008; Rasteiro et al., 2009; Rutten et al., 2012; Wankat, 2002).

Dynamic process simulators have been used as an additional learning tool at HiOA since 2010 (Komulainen and Løvmo, 2014). Our experience shows that simulator training provides industrially relevant practice for large student groups. However, in order to provide prompt assessment of learning outcomes at an individual level, an effective personal feedback and assessment tool is required.

Both industrial and academic experience on simulator training indicate a need for effective automatic assessment measures. The challenge in developing such a tool is to avoid too deterministic measures (i.e. scenario-specific sequences), and to ensure the clarity and measurability of the learning outcomes.

1.2 Introduction to the proposed work

The simulation module is built up using the six categories of the didactic relation model: learning goals, content, learning process, learning conditions, settings, and assessment. These categories are relative to each other i.e. if changes are made in one of the categories this will lead to changes in the other categories (Bjørndal and Lieberg, 1978; Hiim and Hippe, 1998).

Thus, the assessment of the simulation module has to be directly related to the learning goals of the simulation module. In the following, we suggest measuring the theoretical knowledge using key performance indicators (KPI) and to measure practical competencies using operator performance indicators (OPI).

1) Key performance indicators (KPI): The evaluation of the performance of any process is a matter of high

priority, as it is necessary to determine how efficient the process is and whether it is being executed as optimally as possible. In the research of Manca et al. (2012), it is indicated that from the 1980s, the scientific community became aware of the industry's need for performance assessment. Therefore, it was necessary to establish quantitative indicators that could help to measure the production efficiency of a process; these indicators are known as Key Performance Indicators (KPIs).

Key Performance Indicators express the performance of a whole process; they measure the performance of all types of equipment that form a plant and of the entire plant itself (Lindberg et al., 2015). In the industry sector, performance indicators based on human factors are called operator performance indicators (Manca et al., 2012), which, conversely to KPIs require a more complex evaluation due to their implicit human attributes.

2) Operator performance indicators (OPI): Kluge et al. (2009) carried out extensive research on different training methods used for process control simulators. They explain several of the goals of simulator training, some of which are summarized below:

- Lead the trainees to an understanding of physical processes, the overall operation of the plant, and system functionality.
- Start-up and shut-down procedures.
- Procedural knowledge for normal plant operation and the use of checklists.
- Operators should be able to improvise and adapt to the contingencies of abnormal events.

The goals of simulator training are thereby to meet an overall main objective: efficient operator performance.

From the research of (Nazir et al., 2015), several relevant factors can be recognized that can be considered as operator performance indicators. In the process industry, there are two kinds of operators, Control Room Operators (CROPs) and Field Operators (FOPs). One of the most important features of the teamwork between these two kinds of operators is communication. Effective collaboration between CROPs and FOPs leads to the necessary actions to avoid accidents. Therefore, one important OPI is effective communication. Another OPI that can be associated with the teamwork between CROPs and FOPs is the accomplishment of tasks. Process safety is determined by different capabilities that must be associated with operators. Hence, these capabilities are related to OPIs as well: the ability to interpret the available information; ability to identify abnormalities; understanding the process in terms of operation, equipment, and instruments; being able to interact with different teams and deal with abnormal and escalating situations. Another specific characteristic of great importance, which is also related to OPIs, is time. The time taken to execute certain tasks and more specifically, the time taken to deal with abnormal or emergency scenarios, as this is a direct reflection of the responsiveness and attention skills of the operator (Nazir et al., 2015).

Similarly, based on the research conducted by (Nazir et al., 2012) on situation awareness in industrial plants, Manca et al. (2012) identified some characteristics that are related to the concept of OPI. These characteristics are:

- level of knowledge of the fundamentals of the process;
- the role played by the streams involved in the process;
- the ability to run the process under new conditions;
- the ability to deal with abnormal situations;
- the ability to establish a safety culture and
- the ability to coordinate actions.

There is a common factor in the last four studies referred to above, namely the understanding of the process; this can be considered as one of the most important OPIs, as good performance is based on good knowledge of what is done. Kluge et al. (2009) suggested that "knowledge of how to operate the plant to achieve certain goals can lead to good performance". Nevertheless, it is becoming a challenge for operators to obtain good and sufficient knowledge of the processes they operate due to the great advancements in automation, which are more and more complex and lead to information overload and difficulties related to human machine interface (Nazir et al., 2014; Zou et al., 2015).

Nazir et al. (2013) mention the relevant role played by the execution of an appropriate performance evaluation of the operators. The authors suggest that a correct assessment of the operators is also part of a well-designed training method, in order to reduce the number of accidents occurring in the industrial sector and their impact. It is indicated in the study, that the assessment procedure should be completely objective, in order to guarantee consistency, quantitative assessment, repeatability, and neutrality. Therefore, the assessment process must be automatic. In order to do so, the specific characteristics that the system will evaluate must be identified. These are: OPIs, KPIs and help requirement analysis. In their article, they present an example of the methodology of performance assessment for a catalytic inject process and a C3/C4 splitter. The operator performance indicators evaluated in this case were: Reaction time, Identification ability, Self-dependence, Attentiveness, Multitask handling, Voice communication, Identification ability, Recalling ability, and Situation handling.

Within the same context, Manca et al. (2012) conducted research where they indicate the importance of the assessment of the training performance of CROPs. The authors indicate that developing these evaluations represents a challenge, because the assessment is based on performance indicators related to

human beings and therefore on their intrinsic complexity, which leads to subjective evaluations by the instructors. Because of this, it is very important to develop assessment methods based on quantitative values and not just qualitative appreciation, so the assessment can be as unbiased as possible. In the research, they present a hierarchy scheme with different categories and classifications that form the overall CROP mark. The structure is used as a basis for determining the importance and the weighting of each OPI for the operator assessment. Each OPI is assigned a different value according to its place in the hierarchy using the Analytic Hierarchy Process (AHP). The authors suggested this method in order to overcome the drawbacks related to the subjectivity of the trainers.

Characteristics of the OPIs: One of the main features of OPIs is that they are intrinsically related to human factors as they are linked with the assessment of human beings; this is precisely what makes their evaluation so complex. However, Manca et al. (2012) explain that OPIs are not only based on human factors, there are other parameters that also contribute to the OPIs' definition, such as consistency and association.

2 Materials and Methods

2.1 Software tools for simulator training

The dynamic simulation software used is K-Spice® by (Kongsberg, 2016). K-Spice® is a modular simulation tool for oil and gas unit operations based on first principles physics, chemistry, and engineering.

Exercise Manager is an automatic assessment software product for the K-Spice tool. The simulation model used for the study is a generic oil and gas production simulator model that consists of a three-stage, three-phase oil and gas separation train, the utility systems, and emulated control and safety systems. An overview of the plant is given in Figure 1. More details on the model and the assessment tool are given by (Komulainen and Løvmo, 2014).

2.2 Software tools for simulator training

1) Sample selection: All the participating students attend two different courses at HiOA.

2) Data collection: The anonymous data collected included a multiple-choice questionnaire and the numerical results of the final exam. The questionnaire included several questions about simulators as an additional learning tool, and was evaluated on a 5-point scale. The questionnaire was given to the students at the end of the simulation module. The exam results were obtained from the teacher, who prepares and grades the final exam.

3) Data analysis: Questions on whether simulation enhanced the students' learning outcomes were evaluated on a 5-point scale, the percentages for "agree" and "highly agree" are presented in the following. The marks of the simulation task(s) in the final exam were compared to the average marks in the final exam.

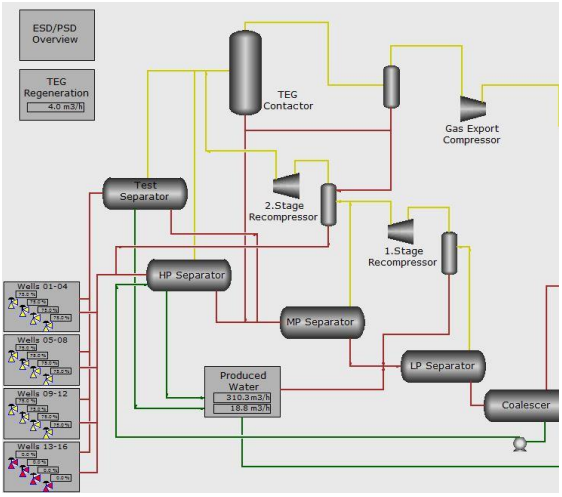


Figure 1. Overview of the large-scale oil and gas production plant model.

3 Teaching and Learning in Simulator Training

3.1 Teaching and learning in simulator training at HiOA

The simulator training at HiOA follows the industrial briefing – simulation – debriefing structure. During the two-hour briefing session, the teacher presents the simulator, the dynamic trends, and the tasks in a classroom for all the students. For the four-hour simulation sessions, the students are divided into larger groups. Typically, the students work on familiarization tasks (60-75min) before the simulation scenarios (2-3h). The students start writing a preliminary simulation report during the simulation session, and spend approximately two hours afterwards to finish the report before the debriefing workshop. In the two-hour debriefing workshop, the students compare and discuss the simulation results in new groups of four students. At the end of the workshop, the teacher facilitates the summarization of the simulation results and of the overall experience on a whiteboard. The total time spent on one simulation training module is 7-10 hours.

The teacher explains the basics of the simulation tasks and gives a simulation demonstration during the introduction lecture. During the simulation sessions, the teacher has an instructor role, only providing help if the student group cannot find the solution themselves. In the workshop, the teacher is a facilitator, setting a framework for the group discussions on the simulation results and guiding the final plenary presentation of the results. The teacher gives the students feedback during the simulation sessions and the workshop, and grades the simulation reports.

The simulation tasks aim to enhance social interaction in small groups while the main focus is for each student to learn by doing the simulation tasks and reporting at their own pace. Discussions on the simulation results are encouraged during the simulation sessions and during the debriefing workshop, i.e. learning from peers and through reflection.

3.2 Current feedback and evaluation methods for simulator training

There is no feedback during the simulation scenarios if the students do not ask the instructor questions. During the debriefing workshop, students get feedback from their peers.

The learning outcomes of the simulation module are measured using the results of the formal final exam.

The students evaluate the simulation module as part of the compulsory report using a multiple-choice questionnaire.

3.3 Experience with simulator modules at HiOA

In the following, the results of two different simulation modules, namely laboratory distillation system and industrial large-scale oil production facility, are presented. The simulation modules were taught to two groups of chemistry students and two groups of electrical engineering students over a period of four years.

The simulation modules were taught in three sessions using briefing–simulation–debriefing (i.e. lecture–computer exercise–workshop) structure, which is typical for industrial simulator training. At the end of the simulation module, the students deliver their simulation reports in groups and present their results in groups at the workshop. The instructor for all simulator modules was the main teacher of the course.

The undergraduate chemical engineering course (fall 2010-spring 2011, 20 chemistry students) where mandatory dynamic distillation simulator exercises were given prior to laboratory experiments: 95% of the chemistry students agreed that simulation enhanced their learning. The average final exam result was 56%, whereas the simulation tasks received an average mark of 70% (Komulainen et al., 2012).

The results for the undergraduate chemical engineering course (fall 2011-spring 2012, 20 chemistry students) were similar, 90% of the students agreed that simulation enhanced their learning. The average final exam result was 43%, whereas the four simulation tasks received an average mark of 47%. The reason for the generally lower exam scores in 2012 was the change of exam type from written to multiple-choice with similar calculation task (Komulainen, 2013).

The undergraduate course in dynamic systems (fall 2013, 60 electrical engineering students) resulted in 97% of students agreeing that simulation exercises increase their understanding of process dynamics in fluid systems. The average final exam result was 59%, whereas the simulation tasks received an average of 48%. One possible explanation for the low score of the simulation tasks was an unclear simulation chart (Komulainen and Løvmo, 2014).

The following year (fall 2014, 60 electrical engineering students) in the exam, the simulation chart was prepared with better resolution and clearer marking of the axes. The final exam result was 58% on average and the simulation task received an average mark of 54%.

In the final exam, the students scored higher than average when the simulation exercise was related to a practical laboratory experiment, and lower than average when the simulation results were not applied afterwards. One possible explanation is that group work without direct feedback might lead to misconceptions.

The students' evaluation of the simulation module and the students' evaluation of their own learning from simulation were very positive for all the groups. The students learn to use industrially relevant tools and their understanding of industrial processes increases.

3.4 Conclusions based on previous experiences

Utilization of industrial large-scale simulators enables students to gain additional skills: industrially relevant process knowledge, and teamwork skills. However, the feedback and assessment system needs to be developed further in order to clearly indicate whether the students have reached the learning goals.

4 Suggested Practices

4.1 Suggested effective assessment method for simulator learning

The main goal of the simulation module is to help the students obtain a better understanding of complex processes and to see the application of theoretical equations and concepts by means of realistic examples and methods. Therefore, there is always an academic commitment to develop revised strategies and procedures that can lead to improvement of the learning outcome.

The aim of this project is to improve the learning outcome of the practices that apply to the simulation module at HiOA. Hence, it is important to be able to measure the knowledge of the students before and after taking the simulation module. This will enable us to make a more formal and reliable evaluation of the benefits of using simulators as a learning tool. In order to achieve this, a diagnostic test based on the required conceptual knowledge about the subject in question should be applied.

The tasks connected to the simulation course have, until now, been based on the students making certain changes to the system and then analyzing the results.

The proposed idea is to add a new section to the simulation module, where the changes in the system will be pre-established, and the students should be able to recognize the abnormal situation and fix it.

The abnormal situation scenarios will be developed using a simulation program associated with the subject or topic of interest. In the case of the present project, which is based on industrial process control, the K-Spice Exercise Manager will be used. The students will have to run different simulation scenarios and observe the possible deviations from normal operations. They will see on the screen the corresponding alarm(s) that will lead them to the source(s) of the abnormal situation. Once the students recognize the problem, they should correct it based on their knowledge of the process. Once the scenario task is completed, a short assessment report will be delivered. The assessment report will be based on strategic performance indicators so that the evaluation is objective and unbiased. The total assessment will correspond to a main performance indicator, which is the Abnormal Situation Management (ASM). This main indicator at the same time may depend on different complementary factors as can be seen from Figure 2.

In Figure 2, the effectivity of the trainee refers to the total time required by the student to fully complete the task. The oil production must be monitored since this is the main goal of the industrial process related to the simulation module, and abnormal situations must be solved as soon as possible and efficiently, in order to avoid major oil production losses. It is also very important to monitor the environmental indicators, such as the flare flow rate or the produced water composition, since abnormal situations can also have serious repercussions for the environment. Another significant factor is the energy efficiency of the process, which is analyzed through the total power consumption of the plant.

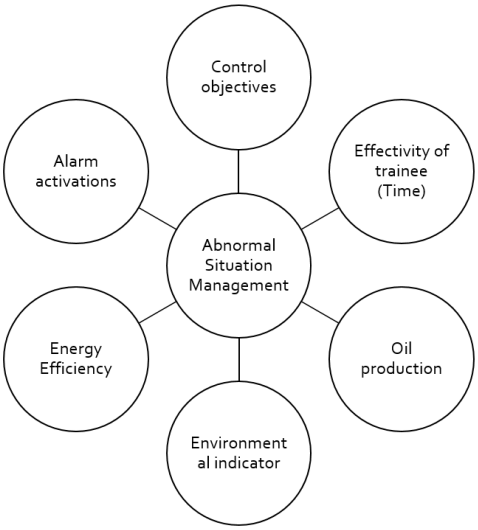


Figure 2. Main performance indicator and complementary factors.

Every abnormal situation in industrial processes is reported by an alarm. The scenarios will be designed such that the problem presented in each task will constantly activate an alarm until the student solves the problem. A record of how long the alarm is active before the problem is solved is indicative of the performance of the student. Finally, the control objectives will be evaluated by the calculation of the integral of the squared error for the controller XC, which indicates how well the problematic controller was tuned, if this is the case. The following equation will be used to determine the total evaluation of the main performance indicator ASM.

$$ASM = \frac{r_{OP} \cdot w_{OP}}{r_{OP,max}} + \sum_i \left(\frac{r_{i,max} - r_i}{r_{i,max} - r_{i,min}} \right) \cdot w_i \quad (1)$$

Where the first term of the equation is related to the oil production (OP), r_{OP}, w_{OP} and $r_{OP,max}$ correspond to the performance measure, the weight of the OP factor and the maximum value of oil production, respectively.

In the second term of the equation, the rest of the factors are evaluated, r_i corresponds to the performance measure of the i_{th} factor, w_i is the weight of the i_{th} factor and $r_{i,max}$ and $r_{i,min}$ are the maximum and minimum value of r_i, respectively.

Each factor makes a different contribution to the total evaluation of the main performance indicator ASM. The Analytic Hierarchy Process (AHP) (Saaty, 2008), was used to calculate the corresponding weight of each factor. This method consists of creating a square matrix based on a pairwise comparison of the factors. The values that indicate how many times one factor is more relevant than the other are according to Saaty's scale. Finally, the matrix entries satisfy the condition $a_{i,j}=1/a_{j,i}$.

Table 1 shows the pairwise comparison matrix for the factors that constitute the main performance indicator. The final priorities associated with each factor (Table 1) correspond to the priority vector of the pairwise comparison matrix, which is the normalized principal eigenvector of the matrix (Brunelli, 2015).

4.2 Specific example of effective evaluation methods for simulator learning

The scenarios must be related to the tasks that the students are going to develop during the first part of the simulation module. The goal is to gradually increase the difficulty of the tasks within the same contexts. In the first part of the module, the students make changes in the system themselves and evaluate the results. In the second part, they are not going to make the changes but to recognize them and solve them.

Table 1. Pairwise comparison matrix for weighing the factors that constitute the main performance indicator.

Pairwise Assessment	ET	OP	EI	EE	AA	CO	Priorities
Effectivity of Trainee (ET)	1	1/4	1/3	1/3	1/2	1/2	0.063
Oil Production (OP)	4	1	1	1	3	3	0.262
Environmental Indicator (EI)	3	1	1	1	3	3	0.251
Energy Efficiency (EE)	3	1	1	1	2	2	0.218
Alarm Activations (AA)	2	1/3	1/3	1/2	1	1/2	0.091
Control Objectives (CO)	2	1/3	1/3	1/2	2	1	0.115

For example, one of the tasks of the first part of the module consists of producing a failure in the level controller of the HP separator by changing the controller to manual mode and decreasing the controller output. As a result, the level in the separator increases and reaches the High-High level, which activates the security alarm, and a partial shutdown occurs. The corresponding assessment scenario will also be based on a controller failure, but the students will not know this in advance. The student will have to run the simulation and observe the system behavior, identify the alarm and solve the problem.

In this particular case, the level will reach the High limit, and it will then stabilize for a moment before reaching the High level again. These kinds of scenarios are also devised with the aim of developing the students' situation awareness, since they must be attentive to recognize the changes in the system.

An example is presented below to demonstrate how to apply the Analytic Hierarchy Process together with (1) to calculate the result of the main performance indicator. The results presented below correspond to a trial test executed by the authors.

As mentioned before, the scenario consists of a failure in the level controller of the HP separator. When the scenario starts, the controller mode switches from auto to manual and the controller output is decreased until the level in the separator reaches the High Level Alarm, then the controller output increases again until the level inside the tank reaches a safe value. This sequence is constantly repeated until the problem is solved, as shown in Figure 3. The solution is simply to switch the controller back to auto. Since no controller tuning is required in this scenario, and the abnormal situation does not affect any environmental aspects of the process, these two factors are not considered in the pairwise comparison matrix developed for the example, which is shown in Table 2.

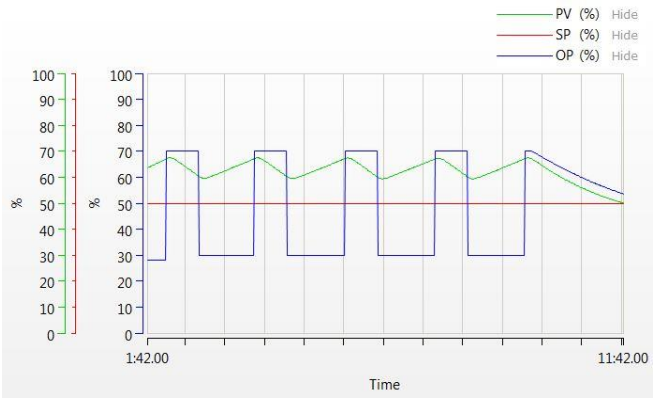


Figure 3. Level controller behavior during the simulation scenario.

Table 2. Pairwise comparison matrix for the example of the level controller failure.

Pairwise Assessment	ET	OP	AA	Priorities
ET	1	1/4	1/2	0.137
OP	4	1	3	0.625
AA	2	1/3	1	0.239

Table 3 shows the values needed for the calculation of each term of (1), and the final calculation of the main performance indicator that correspond to this example. Table 3 also shows the contribution made by each factor to the final value of the Main Performance Indicator. The example was solved in 11.7 min. The minimum time was 5 min and the maximum time was 20 min. There were five alarm activations. In this case, the minimum alarm activations was 2 and the maximum was 10. Finally, the average oil production during the total running period of the example was 908.3 m^3/h and the maximum production under normal circumstances is approximately 980.0 m^3/h. The sum of the values obtained for each factor multiplied correspondingly by their individual contribution gives a final performance of 80%.

Table 3. Calculation of the final value of the main performance indicator.

	r_i	$r_{i,min}$	$r_{i,max}$	w_i	Equation Term
ET [min]	11.7	5.0	20.0	0.137	0.076
AA [-]	5	2	10	0.239	0.149
OP [m^3/h]	908.3	-	980.0	0.625	0.579
	Main Performance Indicator				0.804

5 Conclusions

The simulator training at HiOA currently lacks quick, individual feedback for the participants, and the learning outcomes of the simulator training are not properly assessed after the simulator course. The formal final exam results from HiOA reveal that in spite of the debriefing-workshop after simulator training sessions, some misconceptions remain.

An automatic assessment method is proposed that gives immediate feedback to the students after a scenario is run. The method is based on the evaluation of a main performance indicator that consists of different factors related to the functioning of the process. This main indicator comprises an overall evaluation of the students' progress while dealing with an abnormal situation in the process. The students will receive early and individual feedback on their performance before the workshop, which means they will be able to recognize where there is room for improvement and have the opportunity to work on this before the final exam. Since the instructor will have access to the scenario results of each student, this will also provide the instructor with a clearer picture of how effective the simulator training has been.

The proposed assessment method will be tested at HiOA during the spring and fall semesters of 2017, for the undergraduate courses on chemical engineering and dynamic systems.

Acknowledgements

The authors would like to thank all the participants in the study and the industrial partner Kongsberg Oil & Gas Technologies for providing the software and guidance. Docent Finn Aakre Haugen and Professor Arne Ronny Sannerud are also greatly acknowledged for reviewing the draft paper. Last but not least, thanks to the research funder Oslo and Akershus University College of Applied Sciences, Faculty for Technology, Art and Design.

References

B. Bjørndal and S. Lieberg. In *Nye veier i didaktikken?: En innføring i didaktiske emner og begreper*. Oslo: Aschehoug, 1978.

M. Brunelli. Priority vector and consistency. *Introduction to the Analytic Hierarchy Process*: Springer, 2015

A. Coble, A. Smallbone, A. Bhave, R. Watson, A. Braumann, and M. Kraft. In Delivering authentic experiences for engineering students and professionals through e-labs. *IEEE EDUCON 2010 Conference*, Madrid, Spain, 2010, pages 1085-1090, 2010. doi: 10.1109/EDUCON.2010.5492454

J. E. Corter, S. K. Esche, C. Chassapis, J. Ma, and J. V. Nickerson. Process and learning outcomes from remotely-operated, simulated, and hands-on student laboratories. *Computers & Education*, 57(3):2054-2067, 2011. doi: http://dx.doi.org/10.1016/j.compedu.2011.04.009

T. F. Edgar, B. A. Ogunnaike, and K. R. Muske. A global view of graduate process control education. *Computers & Chemical Engineering*, 30(10–12):1763-1774, 2006. doi: http://dx.doi.org/10.1016/j.compchemeng.2006.05.013

H. Hiim and E. Hippe. In *Læring gjennom opplevelse, forståelse og handling: En studiebok i didaktikk*. 2 ed. Oslo: Universitetsforlaget, 1998.

A. Kluge, J. Sauer, K. Schüler, and D. Burkolter. Designing training for process control simulators: a review of empirical findings and current practices. *Theoretical Issues in Ergonomics Science*, 10(6):489-509, 2009. doi: 10.1080/14639220902982192

T. M. Komulainen. In Integrating commercial process simulators into engineering courses. *IFAC Proceedings Volumes*, 2013, volume 46, pages 274-279. doi: https://doi.org/10.3182/20130828-3-UK-2039.00061

T. M. Komulainen and R. Sannerud. Survey on simulator training in Norwegian oil & gas industry. Oslo and Akershus University College, Oslo, volume 4, 2014, Available: https://skriftserien.hioa.no/index.php/skriftserien/article/view/19

T. M. Komulainen and T. Løvmo. In Large-Scale Training Simulators for Industry and Academia. In A. R. Kolai, K. Sørensen, and M. P. Nielsen, editors, *55th Conference on Simulation and Modelling*, Aalborg, Denmark, 2014, pages 128-137, 2014.

T. M. Komulainen, R. Enemark-Rasmussen, G. Sin, J. P. Fletcher, and D. Cameron. Experiences on dynamic simulation software in chemical engineering education. *Education for Chemical Engineers*, 7(4):e153-e162, 2012. doi: http://dx.doi.org/10.1016/j.ece.2012.07.003

Kongsberg. *K-Spice: A new and powerful dynamic process simulation tool*. Available via https://www.kongsberg.com/en/kongsberg-digital/news/2009/june/0625_kpice/ [accessed August 29, 2016].

C. F. Lindberg, S. Tan, J. Yan, and F. Starfelt. Key Performance Indicators Improve Industrial Performance. *Energy Procedia*, 75:1785-1790, 2015. doi: http://dx.doi.org/10.1016/j.egypro.2015.07.474

D. Manca, S. Nazir, and S. Colombo. Performance Indicators for Training Assessment of Control-Room Operators. *Chemical Engineering Transactions*, 26:285-290, 2012. doi: 10.3303/CET1226048

D. Manca, S. Nazir, F. Lucernoni, and S. Colombo. Performance Indicators for the Assessment of Industrial Operators. In B. Ian David Lockhart and F. Michael, editors, *Computer Aided Chemical Engineering*. Volume 30, pages 1422-1426: Elsevier, 2012. doi: http://dx.doi.org/10.1016/B978-0-444-59520-1.50143-3

C. Martin-Villalba, A. Urquia, and S. Dormido. Object-oriented modelling of virtual-labs for education in chemical process control. *Computers & Chemical Engineering*, 32(12):3176-3186, 2008. doi: http://dx.doi.org/10.1016/j.compchemeng.2008.05.011

S. Nazir, S. Colombo, and D. Manca. The role of Situation Awareness for the Operators of Process Industry. *Chemical Engineering Transactions*, 26:303-308, 2012. doi: 10.3303/CET1226051

S. Nazir, S. Colombo, and D. Manca. Minimizing the risk in the process industry by using a plant simulator: a novel approach. *Chemical Engineering Transactions*, 32:109-114, 2013. doi: 10.3303/ACOS1311028

S. Nazir, A. Kluge, and D. Manca. Automation in Process Industry: Cure or Curse? How can Training Improve Operator's Performance. In P. S. V. Jiří Jaromír Klemeš and L. Peng Yen, editors, *Computer Aided Chemical Engineering*. Volume 33, pages 889-894: Elsevier, 2014. doi: http://dx.doi.org/10.1016/B978-0-444-63456-6.50149-6

S. Nazir, L. J. Sorensen, K. I. Øvergård, and D. Manca. Impact of training methods on Distributed Situation Awareness of industrial operators. *Safety Science*, 73:136–145, 2015. doi: http://dx.doi.org/10.1016/j.ssci.2014.11.015

M. G. Rasteiro, L. Ferreira, J. Teixeira, F. P. Bernardo, M. G. Carvalho *et al.* LABVIRTUAL—A virtual platform to teach chemical processes. *Education for Chemical Engineers*, 4(1):e9-e19, 2009. doi: http://dx.doi.org/10.1016/j.ece.2009.02.001

N. Rutten, W. R. van Joolingen, and J. T. van der Veen. The learning effects of computer simulations in science education. *Computers & Education*, 58(1):136-153, 2012. doi: http://dx.doi.org/10.1016/j.compedu.2011.07.017

T. L. Saaty. Decision making with the analytic hierarchy process. *International journal of services sciences*, 1(1):83–98, 2008.

P. C. Wankat. Integrating the Use of Commercial Simulators into Lecture Courses. *Journal of Engineering Education*, 91(1):19-23, 2002. doi: 10.1002/j.2168-9830.2002.tb00668.x

Y. Zou, L. Zhang, and P. Li. Reliability forecasting for operators' situation assessment in digital nuclear power plant main control room based on dynamic network model. *Safety Science*, 80:163-169, 2015. doi: http://dx.doi.org/10.1016/j.ssci.2015.07.025

Learning Heat Dynamics using Modelling and Simulation

Merja Mäkelä Hannu Sarvelainen Timo Lyytikäinen

Energy Technology, South-Eastern Finland University of Applied Sciences, Finland, www.xamk.fi

Abstract

In the education of energy and power plant engineers, the learning of heat transfer, its dynamics and process control plays an important role. Deeper touch in dynamic process phenomena in our vibrant times may sometimes be a challenge for students and teachers. Modelling and simulation make a continuously increasing tool as a learning method, in various kinds of application fields. In engineering, using pilot or production plants, by modelling the systems, getting results from simulations and comparing the results to real life data, the intended learning outcomes can be achieved in a varying and motivating way. This paper presents a pilot heat exchanger which was mostly constructed by a few students of energy technology and supervised by their teachers. Some basic physical and identified process models of the heat exchanger are introduced, as well as their simulation results. This heat exchanger is widely used in the basic heat transfer and control system studies of higher education. Positive learning and teaching experiences were already achieved in the design and commissioning phase. As a result, the heat exchanger system offers a multifunctional learning environment with modelling and simulation activities in practice-oriented engineer studies.

Keywords: heat transfer, modelling, simulation, process control active learning, collaborative learning

1 Introduction

Learning and teaching of engineering in our times may be rather challenging. Very often also amusing elements are expected. Collaborative learning in teams and practice-oriented experimental learning may help to overcome these challenges. The need of collaborative and practice-oriented learning elements is already referred in (Schadler and Hudson, 2004). Different kinds of active learning activities, such as analyzing, modelling, simulation, visualization, problem-solving, gaming, are emphasized in many learning and teaching instructions, such as in (Centre for teaching Excellence Canada, 2016; Carleton Science Education Center USA, 2016), for example. This paper presents an educational heat exchanger system, some modelling and simulation methods for activating and successful learning and teaching experiences.

The objective educational heat exchanger describes a plate heat exchanger and comprises two water circulations. One water circulation, the hot circulation, includes a boiler unit with an electric heating element, while another circulation, the cold circulation, uses tap water leading it to a removal pipeline. Heat is transferred from the hot circulation to the cold circulation. All pipelines are isolated in order to avoid significant heat losses.

The educational heat exchanger process can be operated using both in the downstream flow and reverse flow and, and thus this system describes many common heat exchangers and heat recovery facilities. The impacts of the circulation flows and heating power can be monitored. The characteristic curve of the pump may be defined. The controllers can be optimized. This rather simple heat transfer system offers a whole set of varying learning objectives (Figure 1).

The heat exchanger process is provided with several temperature measurement sensors with Pt-100 elements: TI-1, TI-2, TI-3, TI-4 and TI-5 being in the hot circulation, and TI-11, TI-12, TI-13, TI-14, TI-15 in the cold circulation (Figure 2). The pipelines have a square-formed cross section profile, and they have a tight contact in the length of 6 meters. The effective heat transfer area totals 0.18 square meters. The dynamics of heat transfer can be experimented and monitored in several ways.

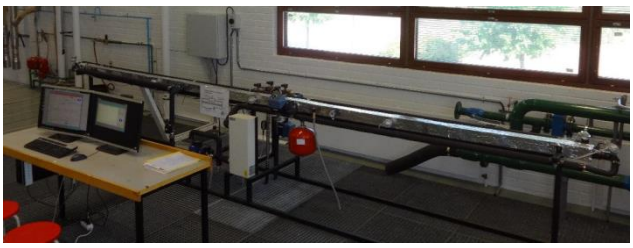


Figure 1. Educational heat exchanger system.

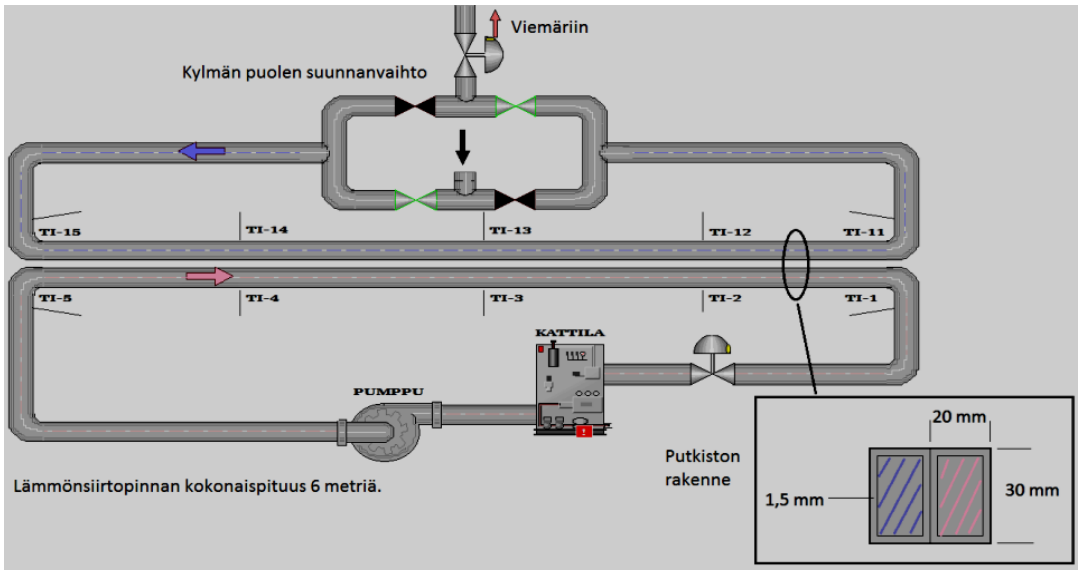


Figure 2. Structure of the educational heat exchanger system.

The temperature of the hot circulation is controlled by the power of the heating elements of the 26 kW flow type boiler (KATTILA), based on the measurement TI-1 (Figure 3). The flow of the hot circulation is controlled by the speed of a feeding pump in the control loop FIC-7. The flow of the cold circulation is controlled by the outlet valve of the control loop FIC-18. The direction of the cold circulation can be changed using a switch (SUUNTA) and related on-off valves, thus enabling the downstream and reverse flow operations.

The educational heat exchanger is provided with a programmable logic controller Siemens Simatic S7-1200 and a PC-based human machine interface (HMI) of Invensys InTouch. The HMI enables users to monitor measurement and control loops in an overview display (Figure 3) and operate them using separate loop window displays. The control system includes sampling of history data for trend curve monitoring, and for numerical tables in case of modelling and simulation.

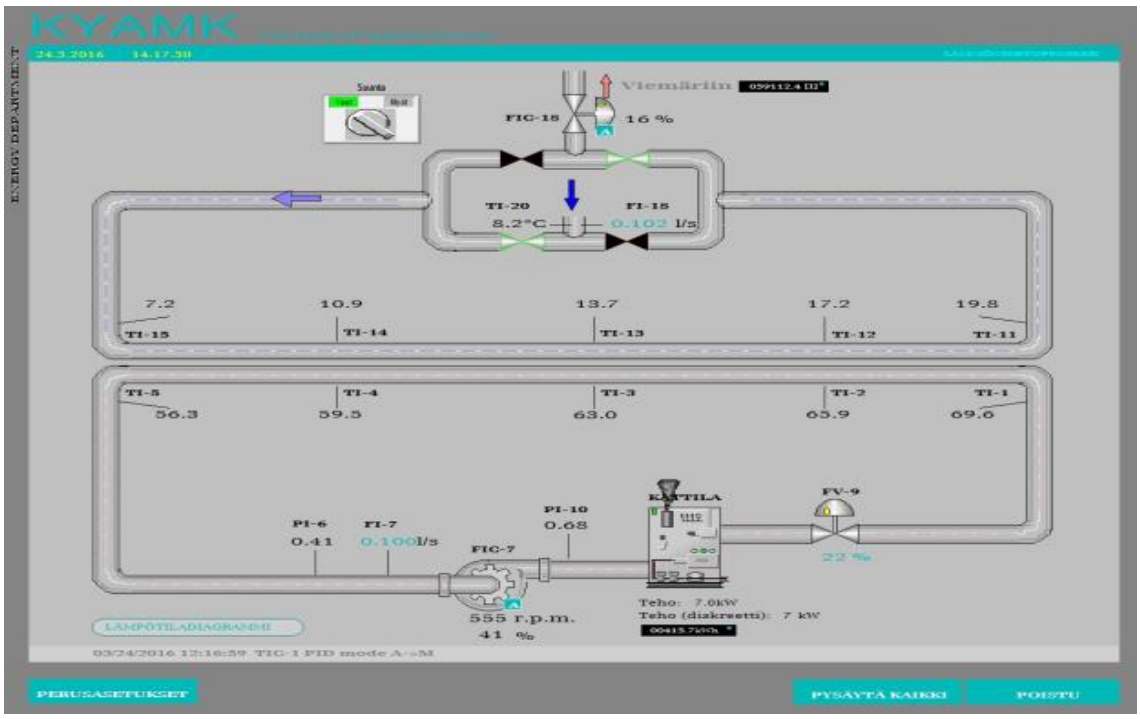


Figure 3. Overview display for monitoring.

2 Physical modelling of heat transfer

In modelling procedures we like to find relationships between interesting, interacting inputs and outputs of systems. Those systems to be modelled can be available process systems, or they may be systems to be designed. Physical models are based on well-known first principles, and they may be applied to rather simple, physically known systems. Model equations which are independent on time, describe steady state models of systems, while differential model equations, for example, present the time-dependent, dynamic behavior of systems. Steady-state models are interesting in process design, while dynamic models help to understand dynamic process phenomena from a control point of view. Modelled systems can be visualized using simulations which are needed in model validations.

We may create static or dynamic energy system models based on flow, mass and enthalpy balances, for example, as described in (Ljung and Glad, 1994). With energy systems we have different aspects. In heat transfer systems one interesting aspect comes from energy efficiency. We often like to know how much energy we do need to make a certain temperature difference in heat transfer.

2.1 First principle model of a heat exchanger

The dynamic behavior of the heat transfer in the educational heat exchanger can be described using a simplified process model of (1). Based on the dynamic heat balance, the change in the energy flow of in the releasing (or receiving) circulation can be given, based on (Ljung and Glad, 1994; Bergman et.al., 2011), as follows:

$$m * c_p * \frac{dT}{dt} = P - k * A * (T - T_n) \qquad (1)$$

m	total water mass, [kg]
c_p	specific heat capacity, [J/(kg K)]
T	water temp. of the input location, [K]
T_n	water temp. of the output loc., [K]
P	incoming heating power, [J/s]
k	average heat tr. coeff., [J/(s*K*m^2]
A	effective heat transfer area, [m^2].

Thus, the trend of temperature in the plate heat exchanger follows an exponential relationship which is dependent on the effective heat exchange area, heat transfer coefficient, mass and enthalpy of water in the circulations. In reality, the specific heat capacity c_p and heat transfer coefficient k are not constants but dependent on the temperature range. The model (1) assumes a constant temperature T_n in the output location, although the temperature T_n also changes due to operation conditions.

2.2 Simulation results based on first principle models

The first principle model of (1) of the educational heat exchanger system was used for the simulation of the temperature T (TI-1).

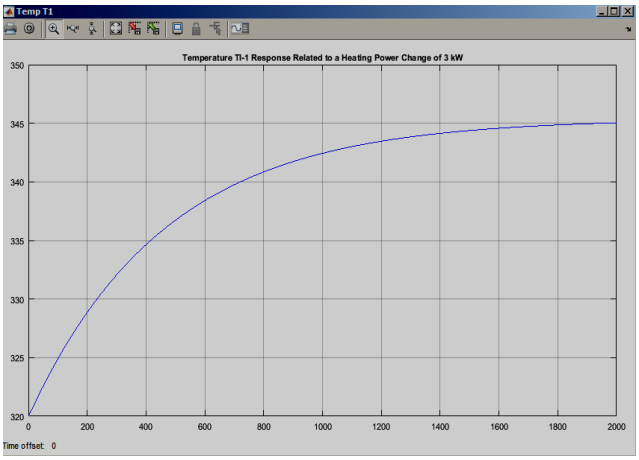


Figure 4. Simulated response of the temperature T (TI-1) related to a heating power change with time [s] in the x-axis and temperature [K] in the Y axis.

The model system was simulated using Matlab Simulink with the following heat exchanger parameters: A=0.18 m^2, T_5=312.05 K, m=10 kg, c_p=4179 J/kgK, k=500 J/sKm^2. The starting temperature T was 320 K. The response in the temperature T (TI-1) results a final increase of 25 degrees in simulations (Figure 4), and respectively in process experiments (Figure 5, red upper curve). The slowness expressed as a time constant, based on (m*cp)/(k*A), gives about 460 seconds. The time delay in the heat transfer could have been added to the simulation model but it was not known, yet.

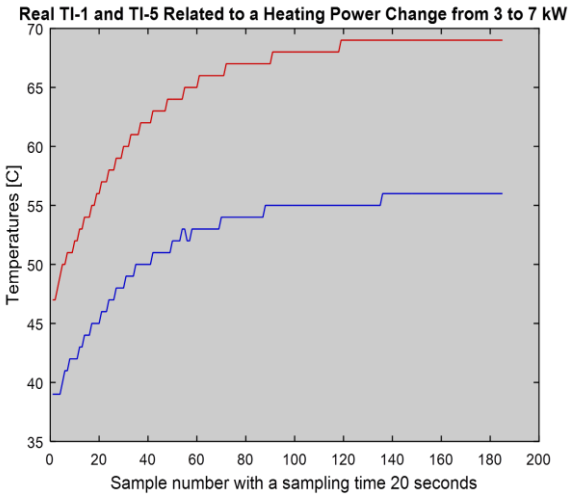


Figure 5. Temperature responses TI-1 (red upper curve), TI-5 (blue lower curve) related to heating power change in a real process experiment.

3 Identification models of heat transfer

Identified models can be achieved based on practical process experiments. Principally, it is a question of fitting sampled data to some ready-made models. These models are often mathematical, or they may even be linguistic such as fuzzy logic models (Yager, 1996; Åström and Hägglund, 1995). On one hand, the data based on simple step response tests can be fitted in time-continuous Laplace models. The most popular Laplace models, due to their physics-related parameters, are the settling first order model and the integrating model. On the other hand, the data sets based on pseudo random binary signal (PRBS) response tests can be fitted in time-discrete AutoRegressive with an eXogenous input (ARX), models, for example, according to (Ljung and Glad, 1994).

3.1 Identified Laplace model with simulation results

The first order Laplace transfer function model between the resulting output Y and excitation input U, for settling dynamics, can be given in the Laplace domain in (2), like presented in (Åström and Hägglund, 1995; Bolton, 2004; Dorf and Bishop, 2004):

$$\frac{Y}{U} = \frac{K}{T*s+1} * e^{-L*s} \qquad (2)$$

Y resulting output
U excitation input
K process gain
T time constant
L time delay
s Laplace operator.

In the first identification modelling case we liked to find out the relationship between the hot circulation temperature TI-1 and the heating power. Using the heating power of 7 kW, the heating process was taken to a steady state. Exact trend curves could be monitored (Figure 6). The heating power was changed to 11 kW, and the temperature T-1 started to increase gradually from 45.4 ºC to 63.4 ºC, after a time delay of 41 seconds. The temperature TI-1 reaches its final temperature 63.4 ºC after about 21 minutes. In the process experiment, the step response of the temperature TI-1 related to the heating power change settles down to a new level and the trend resembles a first order Laplace model response given in (2). The time constant, representing the slowness of the change can be defined to be 570 seconds. Thus, based on the step response curves, and by applying the first order Laplace model of (2), the process model between the temperature TI-1 and heating power can be stated as follows in (3):

$$\frac{Temp}{Heat\ power} = \frac{18/4}{570*s+1} * e^{-41*s}. \qquad (3)$$

The Laplace model of (3) was also constructed in the Matlab Simulink. In a simulation with a heating power change from 7 to 11 kW, the step response in the temperature TI-1 (Figure 6) gives matching results compared to the original process experiment data (Figure 7).

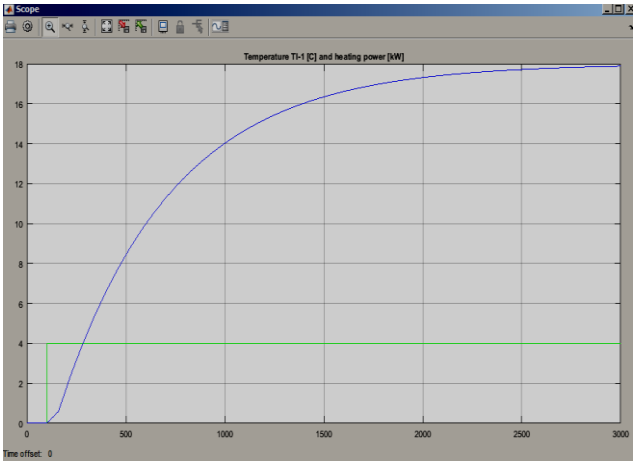


Figure 6. Simulated step response of the temperature TI-1 (blue upper curve) related to the heating power change from 7 to 11 kW (green lower curve).

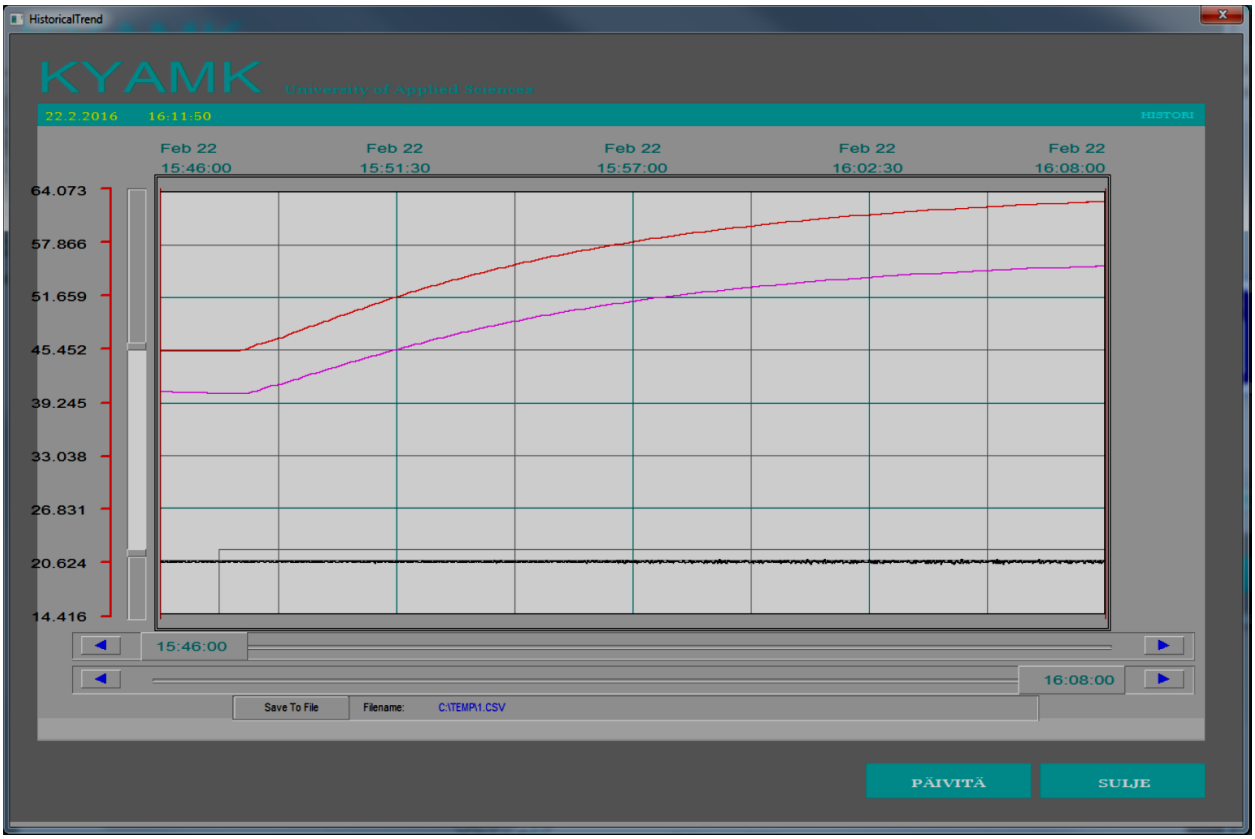


Figure 7. Real step response of the temperature TI-1 (upper red curve), TI-5 (next magenta curve) with the heating power change from 7 to 11 kW (lower stepwise gray curve) in the HMI monitoring.

3.2 ARX model with simulation results

Likewise presented in (Ljung and Glad, 1994; Åström and Wittenmark, 1997) a ready-made ARX model can be applied to practical modelling cases when data sets from process experiments with responses to pseudo random binary signal (PBRS) inputs are available. The ARX model can be given as a Z transfer function, as follows in (4):

$$\frac{Y(z)}{U(z)} = \frac{b_1 z^{-1} + b_2 z^{-2} + \cdots}{1 + a_1 z^{-1} + a_2 + \cdots} \qquad (4)$$

$a_1, a_2, \ldots$ estimated model parameters
$b_1, b_2, \ldots$ estimated model parameters
z Z domain operator.

According to (Ljung and Glad, 1994; Åström and Wittenmark) the parameter estimation of vectors' a and b in (4) can be computed using the least squares method. A process data set, comprising input-output data is fitted to the model structure, and the parameter estimation computing gives the parameters $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$.

The ARX identification model was applied to a data set comprising the temperature TI-1 and heating power. The heating power was set randomly stepwise in 7 and 11 kW giving the response presented (Figure 7). Using the Matlab IDENT for the parameter estimation of an ARX model, the model parameters $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ could be estimated. The parameter estimation was based on a sampled data set with a sample time of 20 seconds. The heating power was switched to 7 and 11 kW, while the temperature TI-1 varied between 63 and 45 Celcius degrees. The first parameter estimation procedure using four a and four b parameters gave rather matching results (Figure 8), but the fitting procedure could be improved.

With eleven a, and ten b parameters the fit could be improved to be about 98 %. The modelling procedure should have been validated using another data set of the process experiment but the validation data was not available. Based on the PRBS process experiment of the pilot heat exchanger, the dynamic ARX model between the temperature TI-1 and heating power can be given using the model (4) where the estimated vectors a and b are:

a = [1.0000 -0.7586 -0.3667 -0.0649 -0.0215 -0.0870 0.1961 0.0999 -0.0810 0.1643 -0.0821]
b = [0 0.2742 0.0806 -0.0142 -0.0552 -0.0790 -0.1086 -0.0569 0.0160 -0.0528 0.0193].

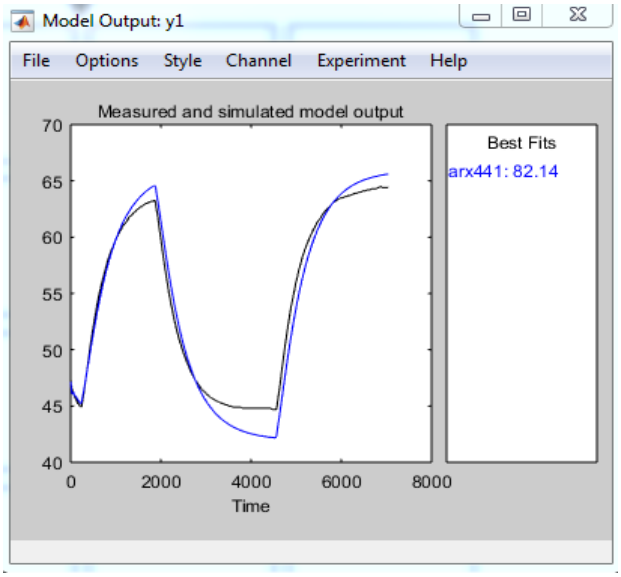


Figure 8. Simulated and ARX model responses, related to PRBS input data.

4 Conclusions

Positive - even attractive and entertaining - learning experiences are expected today also in the education of energy and power plant engineers in many countries. However, hard and some more difficult phenomena should be adapted, as well. Modelling and simulation of practical processes offer interesting learning experiences with many different aspects.

The experiences in learning the dynamics of heat transfer using an educational heat exchanger have been very positive. Combining the theory with practice has helped students to understand and analyze systems with some complexity. Firstly, first principle models based on nature laws could be modelled, constructed and simulated, and finally verified using experimental data. Secondly, based on step response tests, ready-made Laplace models could be parametrized, simulated and compared to the real data. Thirdly, based on PRBS response tests, discrete time series models could be parametrized and simulated.

Cooperative process experiments using the self-made heat exchanger system, modern control system tools, video recordings, photos and shared Moodle course materials with smartphones have been successful elements in the first learning phase. A further interactive analyzing, modelling and simulation phase using Matlab Simulink and Identification Toolbox tools deepened the learning and collaboration. It has been also encouraging to see that students don't have to manage "everything" in Matlab to be able to try basic analyzing, modelling and simulation methods. However, patience is needed in this kind of working. Several practical problems have to be overcome. This kind of a learning process offers a collaborative development environment both to students and teachers and from different disciplines, such as machinery, process technology and system engineering.

In the future, the educational heat exchanger will be provided with remote monitoring and operation in order to support flexible process experiments and data sampling. An interactive learning package based on the Moodle platform with touch-on instructions in plain language will be completed. A special attention will be paid to motivating tools and varying working procedures. Modelling and simulation aspects will be extended step by step. The dependence of specific heat capacity on temperature could be examined and included. Partial models will be collected to make an extended multivariable system model of the educational heat exchanger. The controllers could be provided by additional smart properties. Some parts and activities of this learning package could even be included in a web-based Massive Open Online Course (MOOC).

References

T. L. Bergman, A. S. Lavine, F. P. Incropera, D. P. Dewitt. *Fundamentals of Heat and Mass Transfer*. Seventh Edition, pages 68-78. John Wiley & Sons. 2011.

W. Bolton. *Instrumentation and Control Systems*, pages 226-229. Elsevier. 2004.

Centre for Teaching Excellence Canada. *Active learning activities*. https://uwaterloo.ca/centre-for-teaching-excellence/teaching-resources/teaching-tips/developing-assignments/assignment-design/active-learning-activities. University of Waterloo. 17.6.2016.

R. Dorf, R. Bishop, *Modern Control Systems*. 10th edition, pages 530-534. Prentice Hall. 2004.

L. Ljung, T. Glad. *Modeling of Dynamic Systems*, pages 33-43, 83-105, 120, 231-233. Prentice-Hall. 1994.

J. Mikles, M. Fikar. *Process Modelling, Identification and Control*, pages 13-50. Springer Verlag, 2007.

L. Schadler, J. Hudson. The emergency of Studio Courses – an Example of Interactive Learning, in *Effective Learning and Teaching in Engineering*, pages 156-160. RoutledgeFalmer. 2004.

Science Education Center USA. *What is active learning?* http://serc.carleton.edu/introgeo/gallerywalk/active.html Carleton College. 17.6.2016.

R. R. Yager. A Unified View of Case Based Reasoning and Fuzzy Modeling, in *Fuzzy Logic Foundations and Industrial Applications*, pages 5-20. Kluwer Academic Publishers. 1996.

K. Åström, T. Hägglund. *PID-Controllers: Theory, Design and Tuning*, pages 11-24, 298-304. Instrument Society of America. 1995.

K. J. Åström, B. Wittenmark. *Computer-Controlled Systems*, pages 506-514. Prentice Hall. 1997.

OO Modelling and Control of a Laboratory Crane for the Purpose of Control Education

Borut Zupančič Primož Vintar

Faculty of Electrical Engineering, University of Ljubljana, Slovenia, borut.zupancic@fe.uni-lj.si

Abstract

The paper deals with modelling, simulation and control of a laboratory crane for the purpose of control education. There were many similar activities in the past with realisations in causal modelling e.g. Matlab-Simulink. However we wanted to model and control the set-up also in the OO and multi-domain environment Dymola-Modelica using library components instead of mathematical equations to show all the advantages of such approach. The combination with some causal structures to solve certain problems is also discussed. The model was properly validated with some open and closed loop experiments. These results confirm the applicability of the model and the efficiency of the mentioned approach in modern control engineering courses.

Keywords: control education, modelling and simulation, OO approach, multi-domain approach, model validation

1 Introduction

It is extremely important to use miniature laboratory plants in control education. As such plants are very expensive they usually cover only a part of exercises, the basic part is realised in modelling and simulation environment. There are usually two important areas which have to be covered with education: modelling of real plants and control. Control schemes can be validated in simulation environment or on real plants.

However there are two approaches in modelling and simulation: traditional causal or block oriented or input output modelling which originates in analog simulation, in CSSL standard and is nowadays mostly covered with Matlab-Simulink environment (Moller, 2004; Simulink, 2014). However, more advanced approach is based on acausal, object oriented modelling which represents a multi-domain approach (connection of components from different fields) and giving a possibility of building libraries with reusable components. The most powerful tools are based on Modelica language (Modelica Association, 2010; Fritzson, 2004; OpenModelica, 2012), which is supported with several modelling packages. In our activities Dymola was used (Dymola, 2015). Many experiences with such tools in industrial projects and in education (Zupančič and Sodja, 2013) show, that it is possible to produce complex model in shorter periods while the models are very illustrative as they retain physical structure. However the executable models become very complex, it is usually difficult or impossible detect and solve numerical problems (Sodja, 2012). The practice shows that students are much more motivated when modelling with OO tools. This was evident when our traditional Matlab-Simulink modelling courses were expanded with Dymola-Modelica several years ago.

In our courses AMIRA 600 laboratory set-up (Amira, 2001) for basic and also more advanced courses from modelling and control was used. It enables several subprocesses, one is so called Loading bridge which actually models a crane. The mathematical model and the implementation in SIMULINK is described in (Hančič et al., 2015).

This paper describes our efforts to model the mentioned set-up without numerous mathematical equations. Instead a modeller uses prepared components from standard Modelica library. Our goal was only partly fulfilled because some modelling problems were finally solved with causal components.

2 Description of the loading bridge

AMIRA 600 (Amira, 2001) (see Figure 1) is a convenient laboratory set-up, appropriate for basic and also more advanced courses from modelling and control. It consists of an aluminium frame covered with sheets of plexiglass. The plant has different configurations, one of them is also the bridge crane. The set-up is very convenient for modelling and control. There are different control tasks such as proper positioning of the load in minimal time, small angles of the thread, avoiding obstacles, etc.

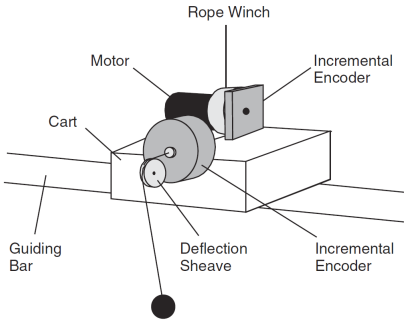


Figure 1. Laboratory set-up for control education: loading bridge.

The laboratory set-up consists of a cart which can be moved along a metal guiding bar by means of a transmission belt. Two proximity switches are mounted close to both ends of the guiding bar. They are used for limiting the position of the cart. The cart carries a rope winch that is used to change the length of the rope. A weight is fixed to another end of the thread. So the length of the thread influences the position of the weight. The lifting and descending of the weight as well as the movement of the cart is realised with two current DC motors enabling the proper positioning of the load (weight). So the DC voltages applied to these motors represent two inputs while the torques produced by both DC motors are proportional to the DC voltages. The outputs are given by three incremental encoders which measure the position of the cart, the length of the thread and the angle between the thread direction and vertical direction (φ). The thread itself together with the load is denoted as a pendulum system. The origin of the coordinate system is the cart in the very left position and the point in which the pendulum is handled to the cart. Figure 2 illustrates the described set-up with some dimensions.

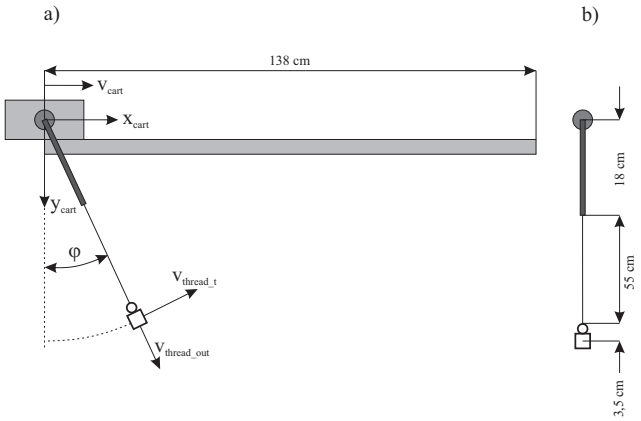


Figure 2. Part a: set-up Amira 600 working as the loading bridge. Part b: pendulum with appropriate dimensions.

Table 1. Physical parameters of the set-up.

Parameters	Values
Mass of the cart+winch	5,7 kg
Mass of the load (weight)	0,143 kg
Length of the guiding bar	1,38 m
Length of the pendulum stick	0,18 m
Distance between the centre of load gravity and the point where the load is fixed to the thread	0,035 m
Maximal length of the rope	0,55 m

Table 1 shows some important plant parameters, which were partly measured or obtained from the documentation (Amira, 2001).

3 Modelling with Modelica

The basic aim of this investigation was to show that Modelica is a convenient approach especially in the modelling of mechanical systems. We intended to use the Modelica standard library and to build an efficient model without usual mathematical modelling with equations. Of course we have to be aware that also a profound knowledge of the mathematical modelling using balance equations in modelling subjects is needed. Even very sophisticated libraries are usually not sufficient the introduction of some changes or developments of customized components in real applications is often needed.

The approach was to build the model with diagram layer (icon or graphical based modelling) and to establish some more efficient solutions using some improvements with textual model layer. Basically the components from two libraries (packages): `Mechanics` and `Blocks` were used.

3.1 Package of mechanical components `Mechanics`

The package `Mechanics` is a part of Standard Modelica library. The library consists of three sub libraries: `Multibody`, `Rotational`, `Translational`.

The library `MultiBody` is a free Modelica package providing 3-dimensional mechanical components to model in a convenient way mechanical systems, such as robots, mechanisms, vehicles. The components - the coordinate system of the world frame, the revolute joints and the rigid bodies have also animation properties.

The libraries `Translational` and `Rotational` are also free Modelica packages which enable modelling of traditional rotational and translational problems which are commonly needed especially in basic modelling courses. The basic components are mass, inertia, spring, damper etc.

It is convenient to model 3D mechanical systems with components from all three libraries, as there are property defined connectors which enable proper conections. As an example we can use two components from the sub library `Mechanics.MultiBody.Joints`, shown in Figure 3. Both components entitled `prismatic` and `revolute` are important in the loading bridge model.

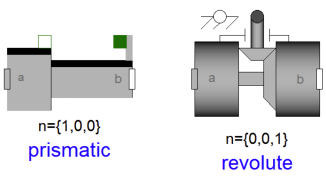


Figure 3. Two components from the sub library `Modelica.Mechanics.Joints`.

The prismatic joint has 1 translational degree-of-freedom. There are two regular connectors (frame `a` and

frame b) which enable to connect the components from the library Multibody. Optionally, two additional 1-dimensional mechanical connections can be driven with elements of the Translational library. This is especially convenient to connect a control force to the 3D mechanical system.

Revolute joint has 1 rotational degree-of-freedom where frame b rotates around axis n which is fixed in frame a. Optionally, two additional 1-dimensional mechanical connections can be driven with a component from Rotational library. This is often used to connect a control torque to 3D mechanical systems.

3.2 Package of causal components Blocks

The library Blocks contains input/output blocks to build up block diagrams. This is actually the implementation of a traditional modelling approach based in causal input/output interactions originating from analog simulation, supported later by a CSSL standard and finally implemented also in Simulink environment. The library is important for the implementation of model parts where the causality is needed, e.g. for the implementation of a control system where inputs and outputs have to be strictly defined.

With components from the library Blocks some new components were implemented and some existing were modified. These components were included into a new library AdditionalComponents, which is a part of the loading bridge model.

3.3 Structure and components of the overall model

Figure 4 shows the hierarchical structure of the loading bridge model. It is implemented with the package CraneModel on the highest hierarchical level. It consists of three sub-packages: AdditionalComponents is the library of components, developed or updated from the existing components from the Standard Modelica library. The second package PilotPlantAMIRA600 is intended to final Amira 600 models. Currently it contains only the model of the loading bridge. The third subpackage Controller is intended to implemented controllers.

3.4 Model of the loading bridge

Figure 5 depicts the top level model of the loading bridge, implemented with model class LoadingBridgeCrane. All components except CartDrive and ThreadDrive are taken from the standard Modelica library Mechanics and model the whole mechanical system with drives. The components CartDrive and ThreadDrive were placed to the library of new and appropriately modified components (AdditionalComponents).

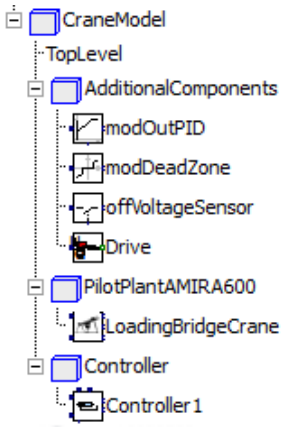


Figure 4. Hierarchical structure of the loading bridge model.

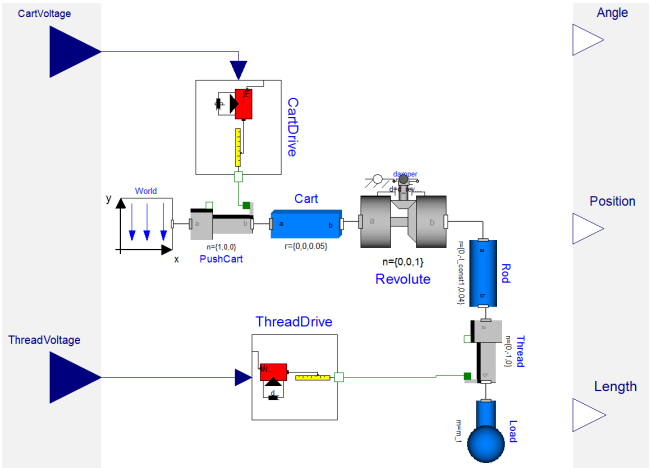


Figure 5. Model of the loading bridge in Dymola-Modelica.

3.4.1 World

Model class World is a part of the standard library Modelica.Mechanics.Multibody. It represents a global coordinate system and the gravity field.

3.4.2 Cart

The cart (model class Cart) is realised with BodyBox component from the Multibody library. The parameters are dimensions, mass, animation features etc. The cart is driven with PushCart element, which is the appropriate actuator for the movement of the cart. The information for the needed movement is calculated in the model class CartDrive.

3.4.3 Pendulum

On the connector b of the Cart the component Revolute is connected. This is a revolute joint which handles the pendulum. The pendulum is built according to the construction shown in Figure 2, b. It is a mathematical pendulum with the whole mass concentrated in a point at the end. The component Rod is a rod with no mass, with appropriate length and it handles the thread which is

realised with the class `Thread`. This prismatic actuator is the implementation of the winch. The movement of the the thread is therefore implemented in the same way as the movement of the cart. The component `ThreadDrive` is almost identical to the component `CartDrive`. So the length of the thread depends on the input DC voltage. The load of the crane is realised with the element `Load` with the whole mass concentrated in the center of gravity point. The distance between the mass point (center of gravity) and the point where the mass is connected to the thread is illustrated in Figure 2, b.

3.4.4 Drive

The model class `Drive` is used in two almost identical classes `CartDrive` and `ThreadDrive`. It transforms the input DC voltage signals into appropriate positions. It consists of causal blocks. The Modelica diagram is shown in Figure 6. This component includes many parameters,

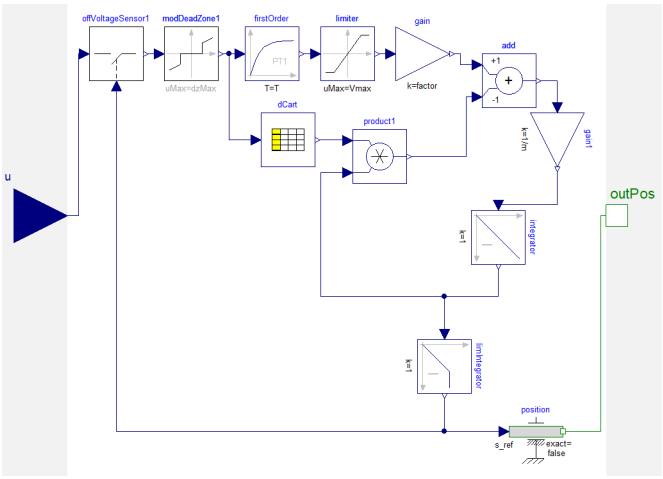


Figure 6. The diagram of model class `Drive` in the Dymola-Modelica environment.

which are needed for proper transformation. The information of the mass of the object and its initial position are also needed. The damping is also important. The experiments show that constant damping is not appropriate. Better results were achieved with the damping which depends on the input voltage. Experimentally obtained curves ware realised with look up tables.

The input to the model is a real variable, representing the DC voltage to the DC motor. This real variable feeds the model class `offVoltageSensor` which models proximity final position switches for limiting the final positions. The output of this class is connected to the input of the component `modDeadZone`, which is a modified standard block `deadZone` from the library `Modelica.Blocks.Nonlinear`. A deadzone in which the voltage is not sufficient for the movement is implemented with this model. The output of the block `modDeadZone` is connected to the lag system `firstOrder` with an appropriate time constant. With this time constant (delay) it was possible to tune the speed of the positioning of the cart and the length of the thread. At the beginning we tried to compensate a fast dynamics with the appropriate damping but it appeared that the appropriate delay, which has to be included, does not depend on the velocity. The output from the lag system goes to the limiter of the input voltage (`limiter`) and is finally in the block `gain` transformed into the appropriate force. The gains 5.17 for the cart and -0.35 for the thread were obtained experimentally. This gains actually represent very simplified models of DC motors. Additionally we improved the model by introducing a damping force. If a force F acts to a mass m, which movement is also damped (damping b), then the acceleration can be evaluated with the equation

$$F - bv = ma \tag{1}$$

where v denotes the velocity and a acceleration. However the model was improved with nonlinear damping where b depends on the DC voltage. The nonlinear function was realised with the look up table `dCart`. Finally the acceleration was obtained by dividing the left hand side of Eq. 1 with the mass m. With two integrations the velocity and the position are obtained (model classes `integrator` and `integrator1`). The calculated position was connected to the position component (`position`) which interacts the causal and acausal modelling parts and implements appropriate movement of the cart or thread.

It is clear that the described modelling is not what was intended at the beginning - to implement fully acausal model, because it introduces partly causal modelling and reduces the efficiency of Modelica language. However this was the most efficient solution of the problem appearing above all in conjunction with the pendulum. Namely we were not able to compensate gravitation and centrifugal forces of the pendulum in open loop experiments what resulted in uncontrolled and unstable movement of the thread. With the solution that the driving actuators obtain the information of the position instead of the force, the compensation forces are automatically generated inside actuators.

4 Controller in Matlab Simulink environment

Dymola-Modelica is an extremely powerful tool for true physical modelling. However for complex experimentations (e.g. optimisation, linearisation, steady state calculation, etc.), for results presentation, e.t.c. it is far from Matlab possibilities. So we decided to use Dymola-Modelica just for the 'physical' part and Matlab-Simulink for all other needs: Simulink for control systems description and Matlab with some Toolboxes for making experiments. We prepared a top level Modelica model which can be used as a Dymola (Modelica) block in the Matlab-Simulink environment. Actually the appropriate connectors which are compatible with other Simulink blocks had to be additionally prepared. Such the top level Modelica model is shown

in Figure 5. Two inputs (cart voltage, thread voltage) and three outputs (cart position, thread length and thread angle) were prepared. Then the Simulink environment to accept Dymola block was properly configured. This block had to be compiled within Simulink before the simulation is started. The control scheme in Simulink is depicted in Figure 7.

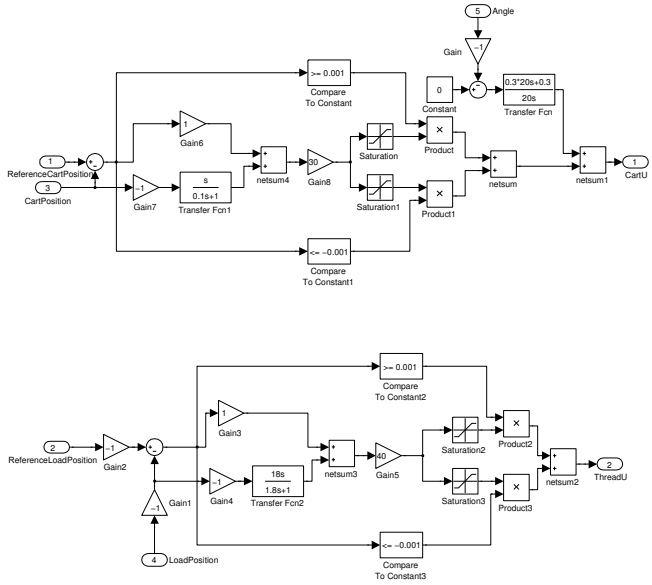


Figure 7. Control system in Simulink.

The upper part is the control scheme for cart positioning realised with PD and PI controllers

$$u_{cart}(t) = K_{Pc}\left(e_{cart}(t) + T_{Dc}\frac{de_{cart}(t)}{dt}\right) \qquad (2)$$

$$u_{angle}(t) = K_{Pa}\left(e_{angle}(t) + \frac{1}{T_{Ia}}\int e_{angle}(t)dt\right) \qquad (3)$$

where u_{cart} and u_{angle} are the appropriate control signals which influence the position of the cart in order to minimize errors between the cart reference position and actual position (e_{cart}) and thread reference and actual angle (e_{angle}). So this part of controller has 3 inputs: the reference value of the cart position (`ReferenceCartPosition`), the actual position (`CartPosition`) and the angle of the pendulum system (`Angle`). So the left part of the structure is described with Eq. 2. PD control was chosen as the process already has the integral behaviour. To realise the real behaviour the output of the PD controller was limited, so that the positive values can change at the interval $1,28V$ - $3V$ and negative values at $-1,8V$ and $-3V$. The behaviour of the model was improved by forcing the output of the controller to $0V$, when the position error was in the range of $\pm 0,001m$. At the right upper part of the scheme the PI control action (Eq.3) was superadded. Namely the influence to the cart position is the only possibility to influence the pendulum angle, which has the reference value 0°.

The lower part is the control scheme (Figure 7) handles the length of the thread. Here the PD controller was used

$$u_{thread}(t) = K_{Pth}\left(e_{thread}(t) + T_{Dth}\frac{de_{thread}(t)}{dt}\right) \qquad (4)$$

where u_{thread} is the appropriate control signal which influences the position of the length of the thread in order to minimize the error between the thread reference position and actual position (e_{thread}). Again the PD controller is used as the process itself has an integral character. The reference length of the thread is in Figure 7 signed with `ReferenceLoadPosition`), and the actual length with *CartPosition*. To match the real behaviour the output of the PD controller is limited, so that the positive values can change at the interval $3V$ - $5V$ and negative values at $-2,8V$ and $-5V$. The behaviour of the model was improved by forcing the output of the controller to $0V$, when the position error was in the range of $\pm 0,001m$.

The parameters of all three controllers are shown in Table 2.

Table 2. Controller parameters

K_{Pc}	T_{Dc}	K_{Pth}	T_{Dth}	K_{Pa}	T_{Ia}
30	1	40	18	0,3	20

Figure 8 depicts the overall Simulink scheme, where the control structure (Figure 7) was realised with a Simulink subsystem `Control system` and the physical part of the laboratory set-up with Modelica model `DymolaBlock`, which has to be compiled before simulation.

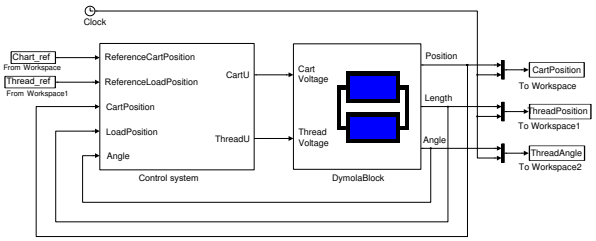


Figure 8. Overall scheme in Simulink with Modelica block.

5 Experiments

It is well known that the model validation is the most important part of each modelling procedure. It is based on comparisons between real measurements and simulation results. Many open loop and closed loop experiments were performed.

5.1 Open loop experiments

The basic validation was performed in open loop with the DC voltage inputs as shown in Figure 9. These inputs

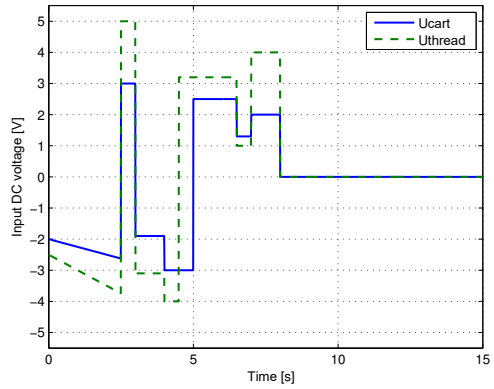


Figure 9. DC motor voltages for open loop experiments.

influenced the real loading bridge set-up and the model. According to several experiments some parameters were tuned and some procedures, which were already commented in the modelling section, were performed. The validation results presented in Figure 10 show that the model behaviour is satisfactory. It appropriately describes the movements of the cart and not so good the lowering and lifting of the weight. Especially it is difficult to tune the angle of the pendulum, where the basic frequency is properly modeled, however real measurements contain also higher frequency components.

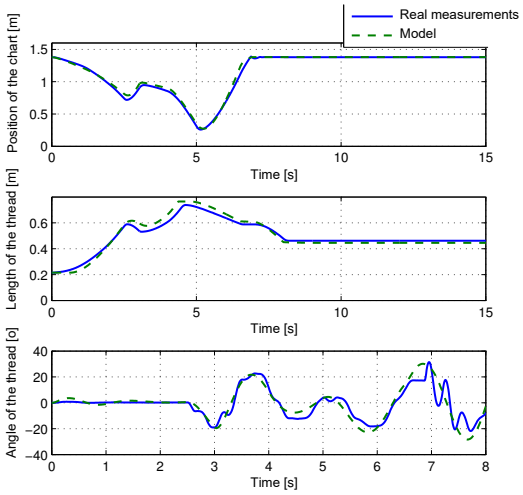


Figure 10. Open loop experiment: comparison of model outputs and corresponding real measurements.

5.2 Closed loop experiments

Figure 11 shows the position of the cart, the length of the thread and the angle of the thread in a closed loop experiment. The same reference signals were applied to the real set-up and the model and the control systems were of course identical. We can notice that the behaviour of the model is reasonable also in the closed loop. The largest deviations are again at the pendulum part, especially with its oscillations.

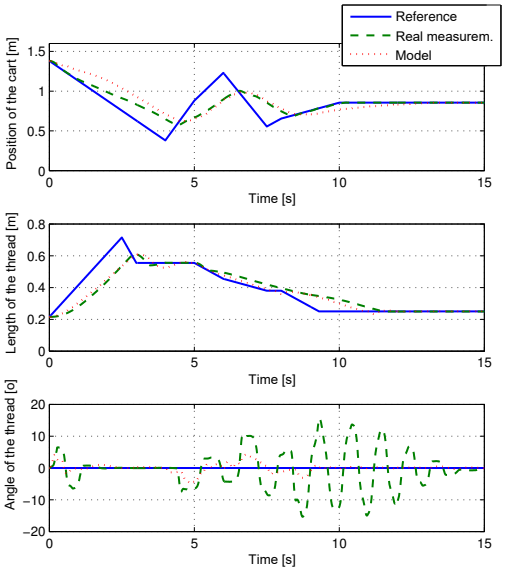


Figure 11. Closed loop experiment: reference signals, model outputs, real measurements.

As the basic goal of the control of the crane is that the load tracks the appropriate trajectory, we also present it in *xy* plane, where *x* and *y* are coordinates of the load. Figure 12 shows the reference trajectory, the trajectory of real measurements and the simulation trajectory.

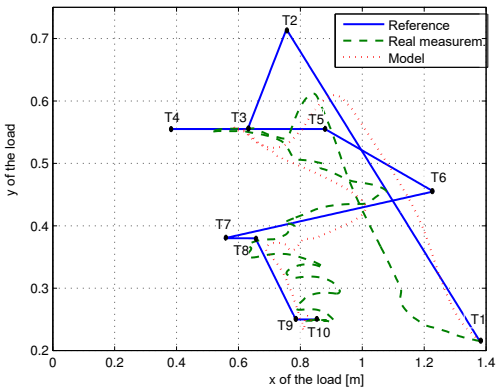


Figure 12. Presentation of trajectories in *xy* plane.

The reference trajectory is defined with points *T1* to *T10*. It has to be mentioned that controller parameters were not strictly optimised as the emphasise was given to the modelling part.

6 Conclusions

The laboratory set-up Loading bridge AMIRA 600 is a very efficient equipment for control education enabling interesting modelling and control courses. The models were previously mostly developed in Matlab-Simulink environment where all balance equations have to be precisely specified. In this investigation we tried to develop an us-

able model without equations, with OO approach using Modelica language and Standard Modelica libraries with mechanical components in Dymola environment. However some difficulties in modelling were later solved with causal approach, which was in Modelica implemented with the Block library showing that the combination of different approaches often gives better solutions. Anyway, such model much better preserves the physical modelling structure, what is important in the education but also in better understanding when model users are not modelling experts.

The experiments results confirm that the model reasonably describes the real system. The worst part of the model is the presentation of the pendulum angle where some higher frequencies also appear in real experiments. Additional efforts will be also devoted to the control system (optimisation of parameters, new control strategies, etc.).

Acknowledgements

The work described in this paper was conducted within InMotion project, co-funded by the Erasmus+ Programme of the European Union, No. 573751-EPP-1-2016-1-DE-EPPKA2-CBHE-JP.

References

Amira. *Documentation Amira PS600 V2.0; Laboratory Experiment Loading Bridge*. Amira GmbH, 2001.

Dymola. *Dymola, Dynamic Modeling Laboratory, User Manual, Volume 1*. Dassault Systèmes, Ideon Science Park, Lund, Sweden, 2015.

P. Fritzson. *Principles of Object Oriented Modeling and Simulation with Modelica 2.1*. IEEE Press, John Wiley & Sons, Inc., USA, 2004.

M. Hančič, G. Karer, and I. Škrjanc. Modeling, simulation and control of a loading bridge. *SNE Simulation Notes Europe*, 25(3–4):165–170, 2015.

Modelica Association. *Modelica Specification, version 3.2*, 2010. `http://www.modelica.org/documents/ModelicaSpec32.pdf`, (accessed May 5, 2016).

C. B. Moller. *Numerical computing with Matlab*. SIAM, Phiadelphia, USA, 2004.

OpenModelica. Open Source Modelica Consortium, 2012. `http://www.openmodelica.org`, (accessed May 5, 2016).

Simulink. *Simulink, Dynamic System Simulation Software, Users manual, R2014a*. Natick, MA, USA, 2014.

A. Sodja. *Object-oriented modelling and simulation analysis of the automatically translated models*. PhD thesis, Faculty of Electrical Engineering, University of Ljubljana, 2012. `http://msc.fe.uni-lj.si/Download//Zupancic/Dissertation_Anton_Sodja.zip`, (accessed May 5, 2016).

B Zupančič and A. Sodja. Computer-aided physical multidomain modelling: some experiences from education and industrial applications. *Simulation modelling practice and theory*, 33(6):45–67, 2013.

A New Approach Teaching Mathematics, Modelling and Simulation

Stefanie Winkler Andreas Körner Felix Breitenecker

Institute for Analysis and Scientific Computing, Vienna University of Technology, 1040 Vienna, Austria
{stefanie.winkler,andreas.koerner,felix.breitenecker}@tuwien.ac.at

Abstract

This paper introduces two different e-Learning environments. Both are used at the Vienna University of Technology to support the courses and exercises in mathematics. There are different level of courses. On the one hand there is a refresher course to support new students who might had some time off before starting their study as well as flatten different school levels of mathematics. On the other hand there are regular mathematical courses in the first two to three semester. Due to improved and advanced possibilities offered by the environment in the last years the system enables the integration of simulation examples. In 2006 the research group Mathematical Modelling and Simulation (MMS) developed an individual web-server to provide students with simulation examples. This server was used in the lectures as well as for practice at home. In the last year also a combination of Moodle and this web application was used to perform tests. This paper should give a short introduction in both systems and compare their advantages and disadvantages. In the outlook a new possibility is presented to combine the advantages of both presented systems.

Keywords: modelling and simulation, education, blended learning, case studies

1 Introduction

In 2004 the research group started to use online tools to improve the quality or at least the administration part of lectures. The first step was using a moodle based platform. The moodle installation was made by an group at the university and is still supervised by the same group. They are as well developing various additional plugins for the moodle based university website. The website is mainly used to organize lecture notes and additional materials. There is also another web application used for lectures dealing with modelling and simulation. This environment has been developed in 2006 and is called Mathematical, Modelling and Tools (MMT). In general it is a simple content management system but with the possibility to connect to simulation environments. This eLearning system is explained in Section 2. (Winkler et al., 2010)

Due to the fact that the research group Mathematical Modelling and Simulation (MMS) was one of the first users of the moodle based university platform, developed the MMT server and won an E-Learning award in 2007 the vice rector for academic affairs commissioned the research group to organize and administrate a refresher course using b-learning. This refresher course was initialised to help students starting their studies subsequently after school. The content of this course is a summary of mathematical knowledge students should have gained at school. The structure and the used eLearning system is explained in Section 2.

2 MMT - Mathematics, Modelling and Tools

2.1 Structure of MMT

The requirements for the website include the possibility to use modelling and simulation examples online. The students should be able to execute a certain simulation example and receive the generated output plot. Another feature is experimenting with these simulations in some kind. The realization enables students to change different parameters of the system and generate the new output to get an idea of the influence these parameters have.

In the last years there where several improvements in the environment. At the beginning the usage of simulation software was restriced to Matlab alone. The current examples show a wider field. There are examples running in Matlab, Simulink, Octave, Anylogic as well as Java. The expansion including Simulink lead to one problem. Compared to all other examples implementations in Simulink need two different m-files to work properly. In case of Matlab and Octave it is enough to simply execute one program file. The Analogic examples use so called Java Applets which represent the interface between Anylogic and webserver.(Körner et al., 2011) In Figure 1 and Figure 2 the structure of the MMT Server and the examples available is depicted. As you can see in Figure 1 on the left hand side all examples are assigned to different topics. In one topic there are more than one implementation. For example different cases of the pendulum model are available. All the pendulum examples are in the folder "'Pendulum"', as shown in Figure 1 marked with number 1. Therefore the theoretical and mathematical basics are explained on the first site of the main topic. The different model implementations explain only the current used method and the different parameters, as marked with num-

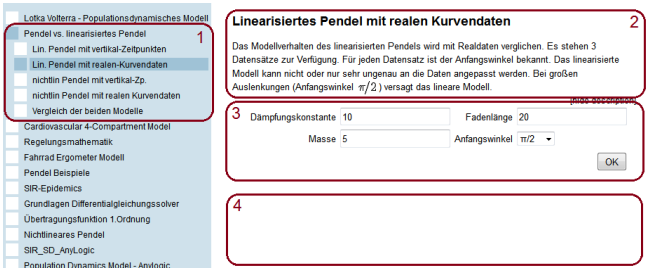


Figure 1. Structure of the MMT system.

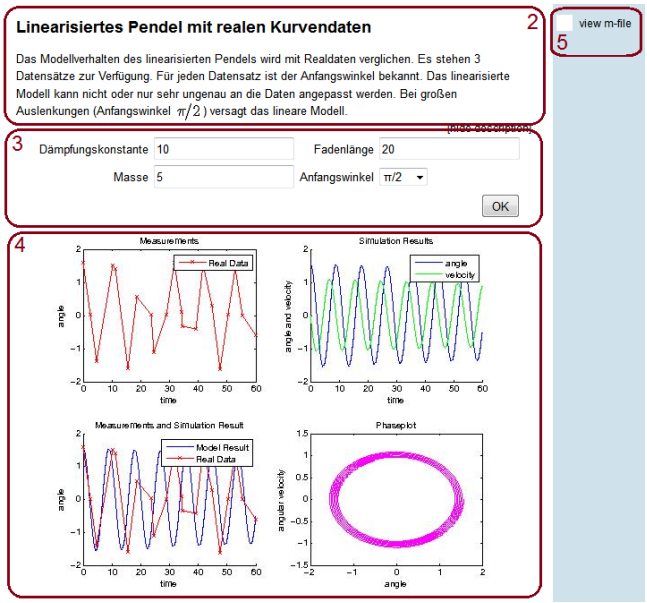


Figure 2. Structure of MMT examples.

ber 2 in Figure 1. In order to know which parameter influences the model output how the students have the possibility to change one or in most examples more than one parameter values. Section 3 shows the parameter area. After setting parameter the student has to summit the parameters by clicking *OK* which leads to the execution of the model. If the example is more complicated it might take some time but usual after some seconds the output plot appears as shown in Figure 1 and 2. Section 5 of Figure 2 offers a link to an m-file. In general this section enables lecturer to upload additional material for the different examples as well. For sure there will be the file containing the corresponding algorithm. It does not matter which simulation environment was used for the implementation. Students have the possibility to download the file and execute or adapt it on their own computer to see how such examples are structured and which commands are used to implement for example the pendulum. (Bicher et al., 2013)

2.2 Usage of MMT

At the moment there are nearly 700 examples stored on the web server. The examples cover various fields of application. Some of the topics are even implemented in different environments to show various realizations or even approaches simulating continuous and discrete models. The examples are used for different courses. The modelling and simulation courses are offered to students of electrical engineering, information technology, mechanical engineering as well as mathematicians. Due to the fact that the knowledge of these groups are not identically the courses are organized differently. Every course contains a certain selection of topics to satisfy the needs of students from different field of studies.

In most of the lectures it is also necessary to pass a test. Up to now these tests are implemented in Moodle. In general the test consists of theoretical and practical questions. The theoretical questions deal with different approaches or even mathematical basic questions. Regarding the practical part the experiments on the MMT server are slightly modified to allow such test questions. In some examples the students have to adapt the parameters of the model to gain certain information of the model behavior, e.g. the state point of a model. Then the results of the practical examples have to be inserted into the Moodle question. Due to the complex structure no randomization in the examples is possible.

3 Maple T.A. - Maple Testing and Assessment

In 2008 the research group was commissioned to initialize a refresher course. There were different requirements the course had to fulfill. The course should be held in the first two weeks of the semester and enables high participation. Additionally the course structure should provide some kind of E-learning support. Due to the fact that the topics of this course are just mathematical subjects the MMT server was not providing a suitable structure. A system called Maple T.A. was chosen.(Urbonaite and Winkler, 2013)

3.1 Structure MTA

Maple T.A. is an interface developed by Maplesoft. In 2008 the software was quite new. It provides an interface to create mathematical questions as well as theoretical questions. The main aspect of the system is the possibility to create practice and test assignments. Therefore it is not only a content management system where different questions are stored but an environment where students can exercise and improve their calculation skills. The administration of examples as well as users is quite easy. In order to create a new question one has to decide which type of question is required. There are different types available, e.g. Multiple Choice, Numeric, Algebraic, graphical questions and so on. For the algorithm the ordinary maple commands are available. The initialization of variables is different but despite of that a user of Maple won't have any problems creating a basic mathematical questions. To give an example a regular if-command in Maple looks like

if a>b then a else b end if .

If one uses Maple T.A. syntax to get the same result it can be written like

$$if(gt(a,b),a,b).$$

For using Maple T.A. to create examples, it is not necessary to know commands of Maple T.A. On the other hand it is helpful to know some of the short devices because they are develop to avoid constant use of the underlying maple kernel, leading to lower capacity requirements. In some cases it is better to use special commands of Maple T.A. Below a short code of a very easy example is given to show the difference to Maple.

$$\$a = range(1,5);$$
$$\$b = range(1,6);$$
$$\$c = range(1,3);$$
$$\$ans = maple(\text{"}if\ \$a > \$b\ and\ \$a > \$c\ then\ \$a\ elif\ \$b > \$a\ and\ \$b > \$c\ then\ \$b\ else\ \$c\ end\ if\text{"});$$

All the created questions can then be combined to assignments associated to a certain topic of the lecture. The assignments can be used during exercises as example resource and of course students can practice examples anytime at home. One advantage of the system is an easy administration of examples and course assignments. Another, even more important benefit, is the usage of a computer algebra system regarding grading of examples. The grading can be easily formalized in a mathematical algorithm ignoring equivalent transformations. On the other hand the question task can be randomized using different parameters in the formulation of even randomize functions, e.g. for calculating a derivative. The variety of the examples increases and students can practice the same assignment more often without repeating exactly the same examples.

3.2 Usage MTA

In the beginning the questions used in the refresher course were very simple. In 2010 the usage of the system was extended. Not only the refresher course but also the basic mathematical courses in the first and second semester were using the system were intense. The level and complexity of the questions increased. This lead to another problem. The grading algorithms had to be improved and get more complex as well. In order to support the acceptance of the system among students the grading provided the possibility to gain partial points if the result is partly right. Two different libraries were established to enable teachers to give partial points if a result is more complicated. On the oterhe hand the library also simplifies the process of creating examples. In Figure 3 an example using the partial grading is shown. This example also uses the second library containing commands to generate randomized matrices with defined properties. The libraries are as well used to guarantee a certain structure of the solution in order to assure similar levels despite randomized factors. The examples cover all topics of the mathematical

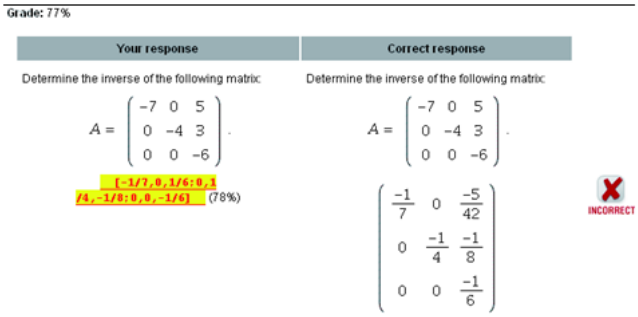


Figure 3. An example working with partial grading.

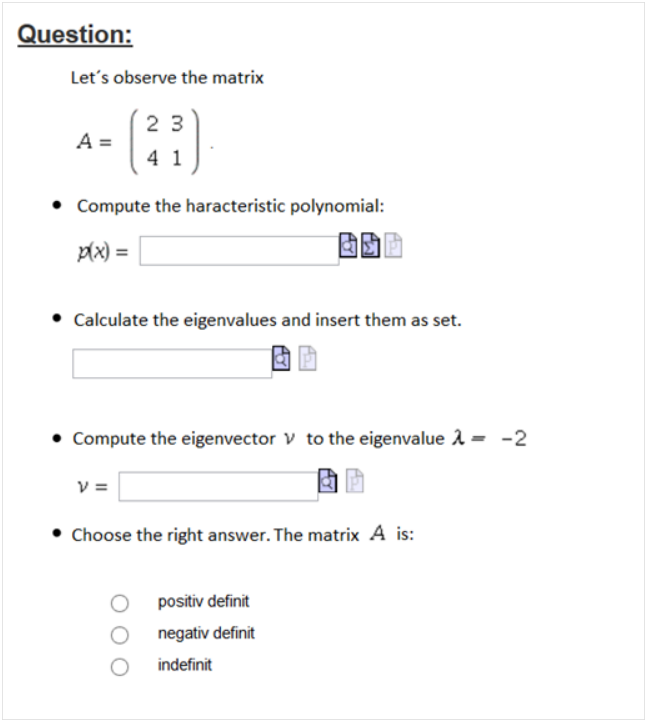


Figure 4. An example used in the written exam.

courses for students of electrical engineering. Examples dealing with Analysis in one and two dimensions, Linear Algebra as well as Vector Analysis and complex analysis are available. Figure 4 shows an example of the chapter Linear Algebra which is also using the libraries mentioned above. Since some years the system is also used to realize exams. Before then the system was used to support the exercises and perform midterm tests. Students have to pass 2 of the 4 provided tests each semester in order to pass the exercise part. For the lecture they have to pass a separate exam. Compared to the small tests during the semester, which only take 30 minutes, the exam lasts 2 hours. This exam is a written test and afterwards a short oral exam follows. Since 2012 the students can decide if they want to do the test as an ordinary written exam or on the computer system. Figure 4 shows a possible question for the exam. The examples are the same but in the online example the variables and functions are changing. Therefore it

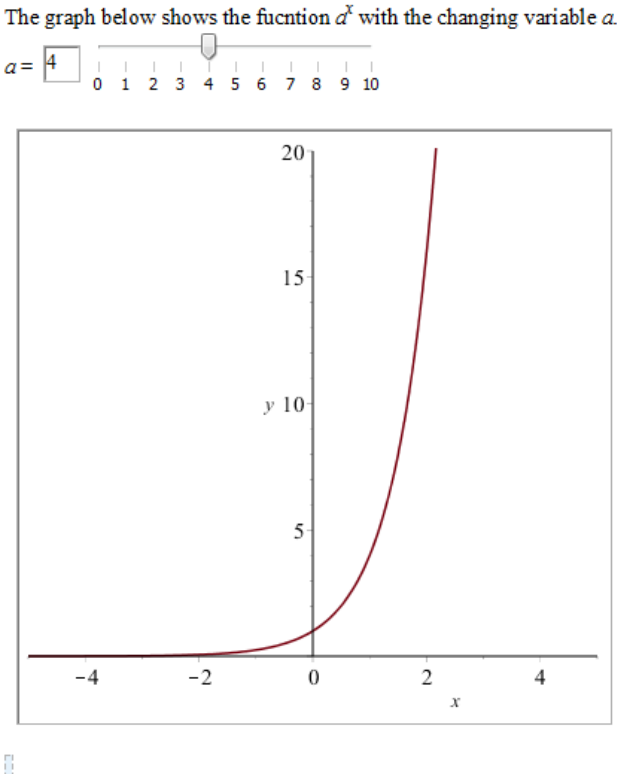


Figure 5. A simple mathapp question used to explain the basics of exponential functions.

is nearly unable to cheat at the examination.

3.3 Mathapps

Due to the fact that the system evolved over the past years there are new interesting features. As mentioned before the development of a test in the web sever MMT is not so easy. A disadvantage is the lack of flexibility regarding randomized test questions. As explained above Maple T.A. enables a wide range of possibilities to randomize parameters and functions of questions. So it would be very feasible to use Maple T.A. for the modelling and simulation courses.

In order to establish simulation examples in Maple T.A. an explanation of Mathapps is necessary. Since some years Maplesoft enables the creation of Mathapps. In general it is a certain form of Maple Worksheets. Before this invention lectures used as well Maple Worksheet to explain mathematical basics benefiting from the algebraic properties of Maple. Compared to an ordinary worksheet Mathapps additionally enable integration of sliders, figures and controllers to design an interactive area as shown in Figure 5. After that an exponential function with parameter a is given. After setting the parameter a using the slider, the graphic is loaded automatically. Such mathapps can be used to explain mathematical principles more easily using the figure of the problem. Regarding modelling and simulation this improvements enables the creation of example similar to MMT examples from above. Students have the possibility to change the parameter and after submitting the new parameter output and figure is actualized automatically.

The implementation using Mathapps can be divided into three different parts. On the one hand there is the so called Startup Code which is executed during loading the question and generates the output plot. In Section 2 it is described that in the beginning there is no output on the MMT server. Students can only see the text describing the example and the parameter area. After submitting the parameter the output appears. In Maple T.A. this can be done automatically. In order to create a efficient example the used algorithm in the Startup Code defines a procedure and therefore can be used again.

If the students submits the parameters using the button *start* the procedure defined in the Startup Code is used to update the output plot. The last part is the grading. There are two different options. On the one hand it is possible to write a grading procedure as well inside the Startup Code. This grading routine then can be used directly in Maple T.A. Another possibility is to formulate the algorithm in Maple T.A. itself. It does not matter which option is chosen. In both cases the current values of all sliders can be used to determine if the chosen input variables students chose are right or wrong.

3.4 MTA - Moodle Connector

In all mentioned cases there is a Moodle course where the lecture notes and materials are organized. A student who enrolls at Vienna University of Technology receives a unique registration number and a password for the university system for course administration. With this account the student can enter the course administration website as well as the moodle based E-Learning website. In previous times the student had to enter the number and password of the university to get an account for Maple T.A. So a student has to go through 3 different websites to get to Maple T.A. Since 2013 it is possible to integrate the Maple T.A. into the e-learning platform of the Vienna University of Technology. This connection is a data interface. If a student registers for the course on the e-learning platform from the university the student is able to get to the examples directly through a link on the course page as shown in Figure 6. There is no need for an additional account in Maple T.A. All results the students gain in the different courses are recorded in Maple T.A. and sent to moodle. This is very useful for administration because also the grade in the end of the semester is created in the moodle system. Therefore the test results can be involved in a formula to generate the over all grade.

4 Case Study

In Section 3.3 the basics of Mathapps created in Maple are explained. Using this tool it is possible to create examples similar to the examples in MMT as shown in Section 2. An implementation of a pendulum example in MMT is shown in Figure 7. Analog Figure 8 presents the realization of the same example inside Maple T.A. using a Mathapp.

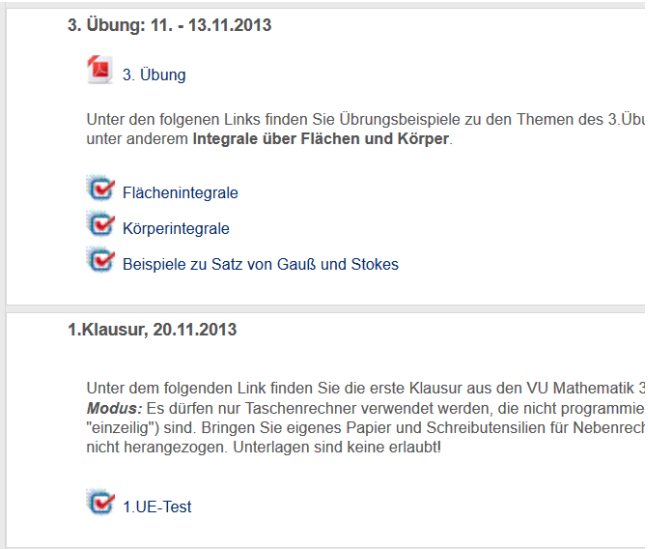


Figure 6. Usage of the moodle Connector to link Maple T.A. assignments.

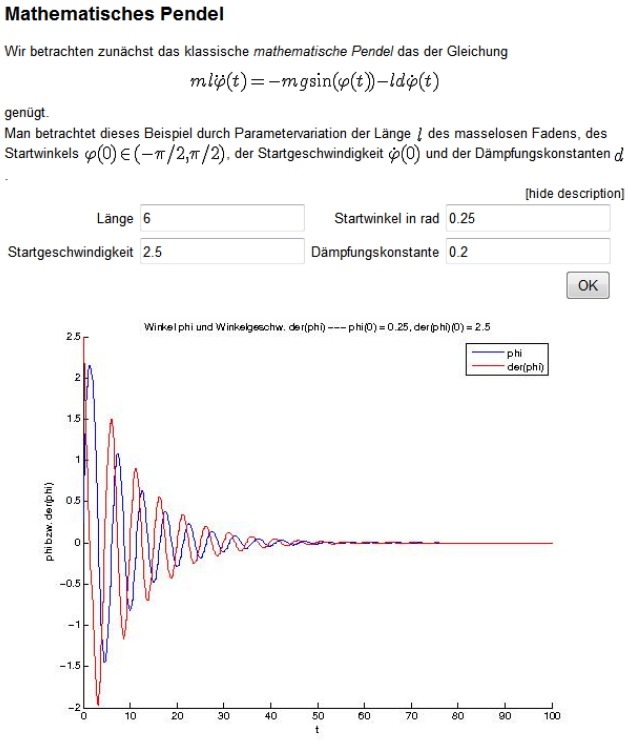


Figure 7. A pendulum example on MMT server.

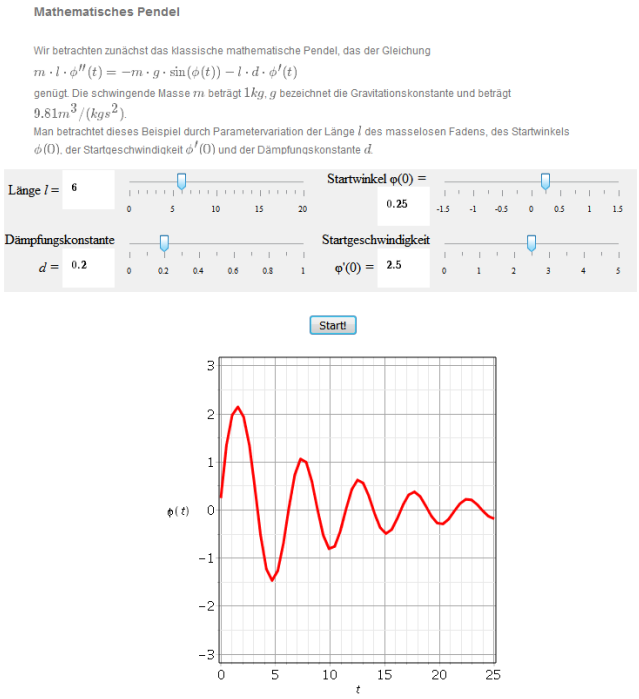


Figure 8. Implementation of a pendulum example in Maple T.A. with Mathapps.

The basic differential equation used to describe the behavior of a pendulum are simple. The implementation of this differential equations is easier done in Matlab or any other numerical environment. It is a greater challenge to implement the same behavior in Maple. In the last years also Maple made some improvements to enable the numerical calculations which are necessary for such Mathapps. Ignoring the fact that the implementation might be easier in Matlab the qualitative behavior of the model is recognizable in both realizations. Both environment allow a default setting by the model. The student can then start experimenting using the sliders to change the parameters or simply insert the parameter value. An advantage of the slider is a restriction of the parameter range. Also in MMT the range of all parameters is restricted but the student would get an error message that a certain parameter value is not possible and then change the parameter in order to execute the example again with a possible value. In case of the Mathapp this can not happen because the student can see the range of the slider. Comparing the two figures the output plot in Maple might be not used to its full potential. Further improvements are possible to enable an easy understanding.

The implementation might be easier using Matlab. The example in Maple T.A. might seem clearer for the students. The area where the parameters can be changed is neatly arranged in the Mathapp implementation. Even the symbols used for the parameter change, if it is a slider a controller or something completely different, can be adjusted to the meaning of the parameter and its variation.

As mentioned in Section 2 there is a test realized using

moodle and MMT in combination. Regarding the lack of randomization in this test as well as level of complexity of questions Mathapps offer more possibilities. One the one hand for grading Mathapps in Maple T.A. every command and library of Maple can be used. Therefore the possible issues which can be covered with the same modelling and simulation question are very variable and different from MMT. Another aspect is the very feasible administration of tests. The acronym T.A. stands for Testing and Assessment therefore the interface provides all setting such tests could need. On the MMT server it was necessary to create a separate question for the test. In Maple T.A. it is possible to use the same Mathapp for practicing and testing. But it is also possible to use the examples in the lecture to explain the basic ideas of the model and the influences of different parameters.

Up to now only some of the modelling and simulation examples are transferred to Maple T.A. Of course it is very easy to formulate all the theoretical qeustions in Maple T.A. For the simulation examples it takes some time to figure out how to implement differential equations efficient in Maple. It is not possible to use the algorithm used in Matlab because the general idea of the two software are completely different.

Despite all the new possibilities of Maple T.A. there are also disadvantages. It is not possible to transfer all the eamples to Maple T.A. because some of the examples are based on Anylogic, Simulink and Java. These Tools are not supported by Maple so we can not use these examples anymore.

5 Conclusions

The comparison of both environment shows that there is no such thing as the best choice. Both systems have their advantages and disadvantages. MMT enables examples using different software products combining them in one content management system. The output on the MMT system shows more detail and potential. But the problem might be that there are much more possibilities in Maple than used in Figure 8. The Mathapps and outputs should be implemented more carefully and then it might be possible to generate outputs as detailed as in Matlab. In Maple T.A. it is not possible to add additional materials. Instead it would be possible to but additional material into the moodle course.

Maple T.A. offers some nice features regarding testing. It is very easy to provide examples in a feasible way for testing and assessing students. Additionally the usage of the moodle Connector mentioned above enables students to connect to the system without even knowing that they changed the platform. At the MMT server every course has its own authorization code. In Maple T.A. practicing and testing is now personalized. The appearance of the examples using sliders and controllers improved. The main advantage of the system might be again the great variety using random variables and function. And as a consequence also the grading offers much more possibilities.

6 Outlook

The conclusion shows that even Maple T.A. which is developed for testing and assessment misses some features. Especially regarding using a combination of different tools. There is an additional environment developed by Maplesoft which could help to erase remaining problems regarding modelling and simulation examples. An aspect which would be nice to extend are teaching possibilities. For Testing Maple T.A. has very useable features. In order to enable more teaching aspects using Maple T.A. Maplesoft came up as well with a solution. The new software is call Möbius (Maplesoft, 2016) and combines teaching and testing aspects. It is a combination of content management system and Maple T.A. but in a more interactive way. It is possible to design something similar to a textbook site including Mathapps as well as Maple T.A. questions.

Imagine a text explaining perhaps the basics of exponential functions. Next an interactive action, e.g. Figure 5, is included. Then there are some examples with its solution step by step. Afterwards an example from Maple T.A. can be integrated. This example actualizes every time. Students can check if they understood the text in order to apply as well the method explained. If the first attempt was wrong there is the possibility to update the integrated examples using the randomized variables in Maple T.A. Using this system students can repeat the lecture in their own tempo and even try examples until the grading is correct.

References

Martin Bicher, Irene Hafner, Andreas Bauer, Carina Pöll, Niki Popper, and Felix Breitenecker. A web-based platform for e-learning and blended learning in modelling and simulation. In *International Conference on Business, Technology and Innovation, in Durres, Albanien, S*, pages 100–109, 2013.

Andreas Körner, Irene Hafner, Martin Bicher, Stefanie Winkler, and Felix Breitenecker. Mmt - a web environment for education in mathematical modelling and simulation. In *Tagungsband Abstracts und Fullpapers*, pages 100–109, 2011. ISBN 978-3-905745-44-3.

Maplesoft. Möbius - online courseware environment that puts stem first!, 2016. `http://www.maplesoft.com/products/mobius`.

Vilma Urbonaite, Stefanie Winkler, and Andreas Körner. Various usage of maple t.a. in mathematics, modelling and simulation. In *ERK - International Electrotechnical and Computer Science Conference*, pages 173 – 176, 2013.

Stefanie Winkler, Andreas Körner, and Irene Hafner. Mmt - a web-based elearning system for mathematics, modelling and simulation using Matlab. In *Proceedings of the 7th Congress on Modelling and Simulation*, pages 1215–1221, 2010. ISBN 978-80-01-04589-3.

Extracting Vibration Severity Time Histories from Epicyclic Gearboxes

Juhani Nissilä Esko Juuso

Control Engineering, Faculty of Technology, University of Oulu, Finland, P.O.Box 4300, FI-90014
{juhani.nissila,esko.juuso}@oulu.fi

Abstract

Monitoring epicyclic gearboxes in vital power transition situations is still a challenge. In this paper, we discuss these challenges with long time vibration measurements through two industrial examples. The first are the two gearboxes in the front axle of a load haul dumper (LHD) from Pyhäsalmi mine and the second a two stage gearbox from Kelukoski water power station (WPS). The LHD was monitored almost continuously for nearly two years until its breakdown. The data from WPS was intermittent from a five month period. We discuss how to find stable conditions for comparable measurements in these cases. For this we utilise a tacho signal from the cardan axle of the LHD and power measurements from the WPS. It is found that in both cases second derivatives of acceleration signals, called snap, respond more quickly to changes in vibration severity. In the LHD case we get clear trends for increasing norms of snap signals. The trends are extracted with nonparametric regression. The shorter measurement period of the WPS makes it impossible to say if its changes are only seasonal. Spectral analysis shows increase in high frequency vibration with time in both cases but provides almost no help for detailed diagnostics.

Keywords: epicyclic gearbox, spectral analysis, higher derivatives, MIT-indices, nonparametric regression

1 Introduction

Typical methods for vibration severity calculations are for example root mean square (rms) values of displacement, velocity or acceleration signals. These may reveal some faults in rotating machines, such as imbalance, but gear and rolling bearing faults often cause high frequency vibration which is more evident in higher derivatives of acceleration. The compact and complex structure of epicyclic gearboxes are no exception. We will present methods to extract vibration severity time histories from epicyclic gearboxes and discuss the difficulties that are encountered.

Signal processing methods are presented in Section 2. These are the Discrete Fourier Transform (DFT) and its inverse (IDFT) for spectral analysis, calculation of derivatives using these transforms, l_p-norms and *MIT*-indices for vibration severity calculations and Nadaraya-Watson nonparametric regression for estimation of *MIT*-trends and other relationships of two variables.

Vibration measurements from the LHD and WPS are described in Section 3 and we will also solve the total revolution times of the gearboxes with the help of some basic number theory. Calculations from vibration measurements are presented in Sections 4 and 5 respectively. Finally the obtained results are discussed in Section 6.

The method for obtaining vibration severity time histories presented in this study is as follows:

1. Find stable and comparable operating conditions of the machine or normalise these conditions computationally.

2. Find which values of norms and derivatives have changed the most during the measurement period and use these for *MIT*-indices.

3. Use Nadaraya-Watson regression to fill in the gaps in measurements. It also extracts the trends more clearly, since there typically still is some variance left in the calculations. Here the scaling parameter is chosen visually.

2 Signal processing

2.1 DFT and derivatives

The vibration measurements are stored as sampled sequences $\mathbf{x} = (x_0, \ldots, x_{N-1})$ of length $T = \Delta t \cdot N$, where Δt is the sampling interval. The spectrum of this sampled signal is calculated with the *Discrete Fourier transform* (DFT)

$$\mathscr{F}\{\mathbf{x}\}_k = X_k = \frac{1}{N}\sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N}. \tag{1}$$

Its *inverse transform* (IDFT) is

$$\mathscr{F}^{-1}\{\mathbf{X}\}_n = x_n = \sum_{k=0}^{N-1} X_k e^{i2\pi kn/N}. \tag{2}$$

Here we have equated the inverse as x_n, because it returns the original signal at the sample points (Briggs and Henson, 1995). The DFT and IDFT pair can be used for differentiation and integration of signals. An algorithm for this consists of calculating the DFT coefficients X_k and

then forming a new sequence $\mathbf{G} = (G_0, \ldots, G_{N-1})$, with $G_0 = 0$ and

$$\begin{aligned} G_k &= \left(\frac{2\pi ki}{T}\right)^z X_k, \qquad 0 < k < N/2 \\ G_{N+k} &= \left(\frac{2\pi ki}{T}\right)^z X_{N+k}, \quad -N/2 < k < 0 \\ G_{N/2} &= \left(\frac{\pi N}{T}\right)^z \cos\left(z\frac{\pi}{2}\right) X_{N/2} \quad \text{(if } N \text{ is even)}, \end{aligned} \tag{3}$$

where z is the order of derivative (or integral when negative). Finally we get the vector $\mathbf{x}^{(z)}$ with the IDFT

$$x_n^{(z)} = \mathscr{F}^{-1}\{\mathbf{G}\}_n. \tag{4}$$

The only problematic part in deriving this algorithm is the term $G_{N/2}$ for even N, which the author has presented in (Nissilä et al, 2014). The algorithm also works with any complex z, in which case we use the principal values of $\left(\frac{2\pi ki}{T}\right)^z$ and $\left(\frac{\pi N}{T}\right)^z$. It then calculates an approximation of the Fourier or Weyl *fractional derivatives and integrals*. This operation is also sometimes called *differintegration*.

The DFT assumes the sequence periodic and thus it is practical to window the signal by multiplying it with a suitable window function to attenuate any discontinuities at the end points of the sequence. We use the window function which was introduced in (Lahdelma and Kotila, 2005)

$$w(t) = \begin{cases} 0, & \text{if } t \le 0 \\ \frac{1}{A}\int_0^t e^{\left(y(y-T/\varepsilon)\right)^{-1}}\,\mathrm{d}y, & \text{if } 0 < t < T/\varepsilon \\ 1, & \text{if } T/\varepsilon \le t \le T/2 \\ w(T-t), & \text{if } t > T/2. \end{cases} \tag{5}$$

Here $A = \int_0^{T/\varepsilon} e^{\left(y(y-T/\varepsilon)\right)^{-1}}\,\mathrm{d}y$ and ε is the portion of T for ascent and descent. Window function w is infinitely differentiable and, therefore, it preserves the continuity properties of the original signal. We use the trapezoidal rule to approximate the integrals in the definition of w.

2.2 l_p-norms and MIT-indices

The *generalised* l_p-norm or *Hölder mean* of vector $\mathbf{x}^{(z)}$ is

$$\left\|\mathbf{x}^{(z)}\right\|_{p,\frac{1}{N}} = \left(\frac{1}{N}\sum_{n=0}^{N-1}\left|x_n^{(z)}\right|^p\right)^{1/p}, \tag{6}$$

for $p \ge 1$. This is the traditional l_p-norm with equal weights $1/N$. The generalisation to cases $-\infty \le p \le \infty$ is done in (Bullen, 2003) with the limiting values

$$\left\|\mathbf{x}^{(z)}\right\|_{p,\frac{1}{N}} = \begin{cases} \left(\sum_{n=0}^{N-1}\frac{1}{N}\left|x_n^{(z)}\right|^p\right)^{1/p} & \text{if } p \in \mathbb{R}\backslash\{0\} \\ \left(\prod_{n=0}^{N-1}\left|x_n^{(z)}\right|\right)^{1/N} & \text{if } p = 0 \\ \max_{n=0,\ldots,N-1}\left|x_n^{(z)}\right| & \text{if } p = \infty \\ \min_{n=0,\ldots,N-1}\left|x_n^{(z)}\right| & \text{if } p = -\infty. \end{cases} \tag{7}$$

The Hölder mean includes many traditional features, such as *minimum value* $p = -\infty$, *harmonic mean* $p = -1$, *geometric mean* $p = 0$, *arithmetic mean* $p = 1$, *root mean square* (rms) $p = 2$ and *maximum value* $p = \infty$.

The dimensionless *MIT-index* for vibration severity evaluation was first presented in (Lahdelma, 1992) and it utilised the rms values of integer order derivatives and integrals. It has been generalised to real order differintegrals (Lahdelma and Juuso, 2011) as

$${}^{\tau}MIT^{p_1,p_2,\ldots,p_M}_{\alpha_1,\alpha_2,\ldots,\alpha_M} = \frac{1}{M}\sum_{m=1}^{M} b_{\alpha_m}\frac{\left\|\mathbf{x}^{(\alpha_m)}\right\|_{p_m,\frac{1}{N}}}{\left\|\mathbf{x}^{(\alpha_m)}_{\mathrm{ref}}\right\|_{p_m,\frac{1}{N}}}, \tag{8}$$

where $\alpha_m, p_m \in \mathbb{R}$, $\sum_{m=1}^{M} b_{\alpha_m} = 1$, τ is the length of the signal and $\mathbf{x}_{\mathrm{ref}}$ is a reference signal from the machine in good condition or low stress. Typically *MIT* increases together with decreasing machine condition. The *MIT*-index can also compare stress levels of different operating conditions.

2.3 Nadaraya-Watson nonparametric regression

Nonparametric regression of two variables using a kernel function was proposed in 1964 independently by Nadaraya and Watson (Nadaraya, 1964; Watson, 1964). Suppose that we have measured values $\boldsymbol{x}$ at the points $\boldsymbol{t}$. Then the estimated value x at t is

$$x(t) = \frac{\sum_{n=0}^{N-1} K_h(t-t_n)x_n}{\sum_{n=0}^{N-1} K_h(t-t_n)}, \tag{9}$$

where K_h is some non-negative and even kernel function for which $\int_{-\infty}^{\infty} K(t)\,\mathrm{d}t = 1$ and h is a scaling parameter so that the scaled kernel is

$$K_h(t) = \frac{1}{h}K\left(\frac{t}{h}\right). \tag{10}$$

Smaller h gives an estimate which follows individual measurements more closely whereas bigger h gives a smoother and more slowly changing function. The kernel in all the regression calculations in this study is a Gaussian (normal distribution)

$$K(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}. \tag{11}$$

3 Measurements and gearbox properties

3.1 Load haul dumper front axle

The measurement setup consisted of four SKF CMPT 2310 accelerometers which were mounted externally onto the LHD's front axle housing and a tachometer on the drive shaft. Accelerometers were located near the planetary gearboxes on either side of the axle and were positioned horizontally and vertically. Measurements were

recorded with a National Instruments CompactRIO 9024 data logger into a solid-state drive (SSD) as binary files of one minute length. Sampling frequency is 12800 Hz, and a built-in antialising filter guarantees that there are no aliases at frequencies that are less than $0.45 \cdot 12800$ Hz $= 5760$ Hz. More information on the measurements can be found from (Laukka et al, 2016).

The measurement points are called right vertical (RV), left vertical (LV), right horizontal (RH) and left horizontal (LH). During the first month of the measurements, the accelerometer cables of LV and LH broke down and were replaced. Two SSDs also broke down almost simultaneously after six months of service, which stopped the whole measurement for a month and a half. A third accelerometer at RH broke down during this stoppage and was replaced. Finally six months before the end of measurements, the accelerometer at RH was broken and after that also the accelerometer at RV. There was only one spare accelerometer left which was installed to LV and the accelerometer at RH was moved to RV (since at this point it was clear that the vertical measurements were more sensitive to the deteriorating condition of the axle.)

At the beginning, measurements were always recorded when the LHD was operating. After the stoppage caused by the broken SSDs the program was modified to record only two hours of data after the LHD starts up.

The cardan axle transfers the power first to a differential in the front axle which has a driving pinion with 9 teeth and a crown wheel with 46 teeth (spiral bewel gears). Based only on some pictures, the actual differential operation is carried out with straight bevel gears that probably have 20 teeth. These do not affect the output ratio if the LHD is not turning. Shafts in the front axle then rotate the sun gears in the epicyclic gearboxes on either side. The epicyclic gearboxes consist of a stationary ring gear (104 teeth), three planetary gears (39 teeth) and a sun gear (19 teeth). These are simple spur gears. Unfortunately, we have not yet received enough detailed information about the bearings in the system other than that they are of tapered roller type. The planet carrier provides the output to the front wheel. This is an example of an epicyclic gearbox in the planetary configuration and driven in reduction mode. Assuming then that the rotational frequency of the drive shaft is $\nu_{driveDIF} = 13.5$ Hz, we get the differential gear mesh frequency

$$\nu_{meshDIF} = 9 \cdot \nu_{driveDIF} = 121,5 \text{ Hz}.$$

If the LHD is not turning, the differential provides the same rotational frequencies to both sides of the axle and the frequencies of the epicyclic gears are (Vicuña, 2010; Immonen et al, 2012) (negative sign means opposite direction)

$$\nu_{sunLHD} = \frac{9}{46}\nu_{driveDIF} \approx 2,64 \text{ Hz},$$

$$\nu_{carrierLHD} = \frac{19}{19+104}\nu_{sunLHD} = \frac{19}{123}\nu_{sunLHD} \approx 0.408 \text{ Hz},$$

$$\nu_{planetsLHD} = -\frac{104-39}{39}\nu_{carrierLHD} = -\frac{5}{3}\nu_{carrierLHD} = -\frac{95}{369}\nu_{sunLHD} \approx -0.680 \text{ Hz},$$

and the planetary gear mesh frequency

$$\nu_{meshLHD} = 104 \cdot \nu_{carrierLHD} \approx 42.43 \text{ Hz}.$$

Gear tooth numbers are typically *relative primes*, i.e. their *greatest common divisor* is 1. This means that it takes a long time for a gearbox to mesh through all of its tooth pairs. To calculate this time for a full revolution of meshes, we seek whole number solutions for the number of revolutions N_{sunLHD}, $N_{carrierLHD}$ and $N_{planetsLHD}$. This leads to *congruence equations*

$$19 \cdot N_{sunLHD} = 0 \mod 123,$$

$$95 \cdot N_{sunLHD} = 0 \mod 369.$$

The solution method can be found in any basic number theory book (Strayer, 1994), and they can also be solved with symbolic mathematical software (such as Mathematica) and even solvers for web browsers exist. The solutions of N_{sunLHD} are $123m$ and $369m$ respectively for all $m \in \mathbb{Z}$. The smallest combined solution is the *least common multiple* of 123 and 369, which is $\text{lcm}(123,369) = 369$. Thus after 369 revolutions of the sun gear, every tooth has returned to their original position and this takes $369 \cdot 1/\nu_{sunLHD} \approx 139.70$ seconds. For most situations this means far too long a signal to analyse (if we can even find that long signals with relatively constant speed). So in practise, we analyse signals whose lengths are at least the revolution times of all the individual components.

3.2 Water power station gearboxes

The water power station at Kelukoski has a two stage epicyclic gearbox. The first (slower) is called gearbox 1 and the second (faster) will be called gearbox 2. Both were monitored with one WBS CM301 sensor (acceleration data was recorded) with sampling frequency 5000 Hz. Every 15 minutes a signal of length 7 s was recorded from both measurement points as WAV files. There were four continuous periods of data collection.

Since the WPS is connected to the Finnish power grid, its output frequency is kept at 12,5 Hz to a high precision (the frequency of the power grid is four times this, i.e. 50 Hz) Thus we can calculate the characteristic frequencies backwards starting from the output of the faster gearbox. Gearbox 2 is in the star configuration, meaning that it has a stationary planet carrier with six planet gears (25 teeth). Output is provided via the sun gear (36 teeth), $\nu_{sunWPS2} = 12,5$ Hz, and input from the gearbox 1 via the

ring gear (86 teeth). The gear teeth are double helical and the gearboxes have plain bearings. The formulas for the frequencies of these components are particularly easy in the star configuration (Vicuña, 2010)

$$\nu_{ringWPS2} = -\frac{36}{86}\nu_{sunWPS2} \approx -5.23\,\text{Hz},$$

$$\nu_{planetsWPS2} = \frac{86}{25}\nu_{ringWPS2} = -18.00\,\text{Hz},$$

and the mesh frequency $\nu_{meshWPS2} = 36 \cdot \nu_{sunWPS2} = 450.00\,\text{Hz}$. The full revolution time calculation leads to congruence equations

$$86 \cdot N_{ringWPS2} = 0 \mod 36,$$

$$86 \cdot N_{ringWPS2} = 0 \mod 25,$$

whose solutions are $18m$ and $25m$ respectively for all $m \in \mathbb{Z}$. The smallest combined solution is $\text{lcm}(18,25) = 450$ and thus 450 revolutions of the ring gear takes $450 \cdot 1/\nu_{ringWPS2} = 86.00$ seconds.

Gearbox 1 is in planetary configuration, but in contrast to the planetary gearboxes in the LHD, it is driven in the other direction (to increase the rotational speed) and the sun gear (31 teeth, output) has more teeth than the planetary gears (25 teeth). The stationary ring gear has 81 teeth. We have $\nu_{sunWPS1} = \nu_{ringWPS2}$ (this middle part of the two gearboxes is a floating installation) and the other frequencies are

$$\nu_{carrierWPS1} = \frac{31}{31+81}\nu_{sunWPS1} = \frac{31}{112}\nu_{sunWPS1} \approx -1.45\,\text{Hz},$$

$$\nu_{planetsWPS1} = -\frac{81-25}{25}\nu_{carrierWPS1} = -\frac{56}{25}\nu_{carrierWPS1} \approx 3.24\,\text{Hz},$$

and finally the mesh frequency

$$\nu_{meshWPS1} = 81 * |\nu_{carrierWPS1}| \approx 117.31\,\text{Hz}.$$

Again, to calculate the time for a full revolution we get

$$112 \cdot N_{carrierWPS1} = 0 \mod 31,$$

$$56 \cdot N_{carrierWPS1} = 0 \mod 25,$$

whose solutions are $31m$ and $25m$ respectively for all $m \in \mathbb{Z}$ and $\text{lcm}(31,25) = 775$. This takes $775 \cdot 1/\nu_{carrierWPS1} \approx 535.11$ seconds.

Before the previous breakdown of gearbox 1 several years ago, it had a slightly different set of gears with 35 teeth in the sun gear, 27 in the planetary gears and 91 in the ring gear. These give the same ratio for the output of the gearbox, but the frequency of the planetary gears was 3.45 Hz and most importantly the mesh frequency was 131.80 Hz. Unfortunately, these older tooth numbers were still thought to be valid for quite some time after the breakdown probably due to some problems with communication. Obviously the lack of such mesh frequency and the the appearance of the actual mesh frequency 117.31 Hz in the spectra was a puzzle in the analysis of vibration measurements from this gearbox. These erroneous calculations and hardly satisfying explanations of their results ended up into some publications (Immonen et al, 2012; Nikula et al, 2015). It seems that the new gearbox is an improved design, since with these older gear tooth numbers we get full revolution of the gearbox in only 135 revolutions of the carrier (which happens due to one of the congruence equations having solutions which repeat very often), which is just 93.21 seconds, considerable less than the 535.11 s of the new gearbox.

4 Calculations from the LHD measurements

For calculations signals were selected from the beginning of most measurement days, when the rotational frequency of the drive shaft was approximately 13.5 Hz. At first this was done manually but after a while with the help of an algorithm which searched for signals with constant enough tacho pulse separations to indicate the desired speed. All of these were still visually checked to select the signals for calculations. Only signals from the beginning of the workday or right after the LHD had been in its weekly maintenance were selected, because then we knew that its bucket was empty and thus the load on the axle was consistent. One could also use the tachometer signal for order tracking the signal, i.e. interpolating the signal to exact revolutions of the cardan axle. This would probably reduce variance in the calculations and make spectral analysis more exact, but because the tacho pulse was recorded badly at times, it would not have been simple to implement. From each signal a 4 s sample was multiplied with the window function (5) using $\varepsilon = 10$ and this new signal was differintegrated with the algorithm (4). Band-pass filtering was performed with an ideal filter at cut-off frequencies 3 Hz and 5000 Hz. From each end of the signal 20% was rejected and the remaining 2.4 second signal (approximately the time it takes for the carrier to rotate once) was used in the calculation of generalised l_p-norms. All the calculations were performed with Matlab.

Figure 1 shows the trend of values ${}^{2.4}\boldsymbol{MIT}_4^2$ from the point RV and a fitted regression estimate with $h = 20$. In (Nissilä et al, 2014) it was already demonstrated that the relative increase of norms of snap signals were bigger than norms of acceleration signals and that the order of norm had very little effect. The fitted regression curve is almost an increasing function with smooth steps. This could mean several different faults or faults which worsen with time.

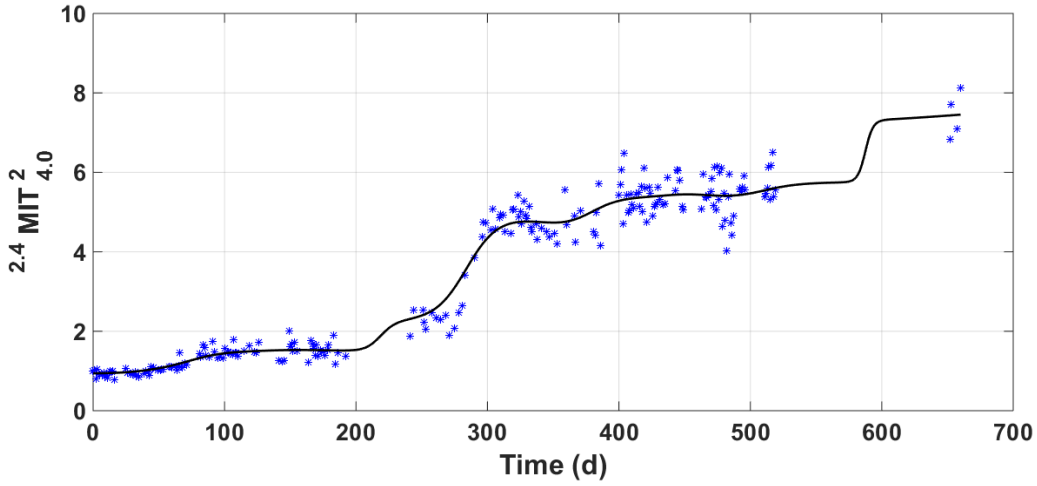


Figure 1. Trend of ${}^{2.4}\boldsymbol{MIT}_4^2$ from the point RV in the frequency range 3 - 5000 Hz and regression estimate with $h = 20$

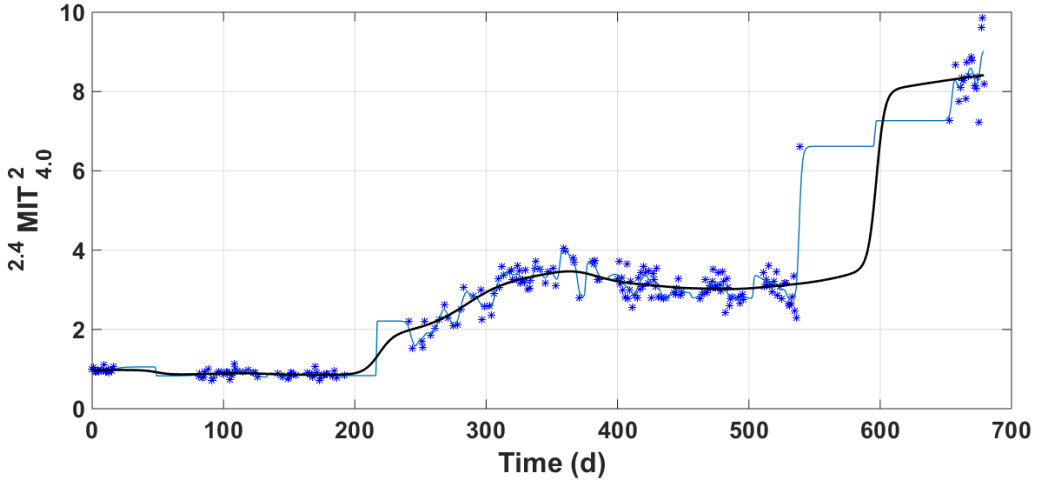


Figure 2. Trend of ${}^{2.4}\boldsymbol{MIT}_4^2$ from the point LV in the frequency range 3 - 5000 Hz and regression estimates with $h = 20$ (black) and $h = 2$ (blue)

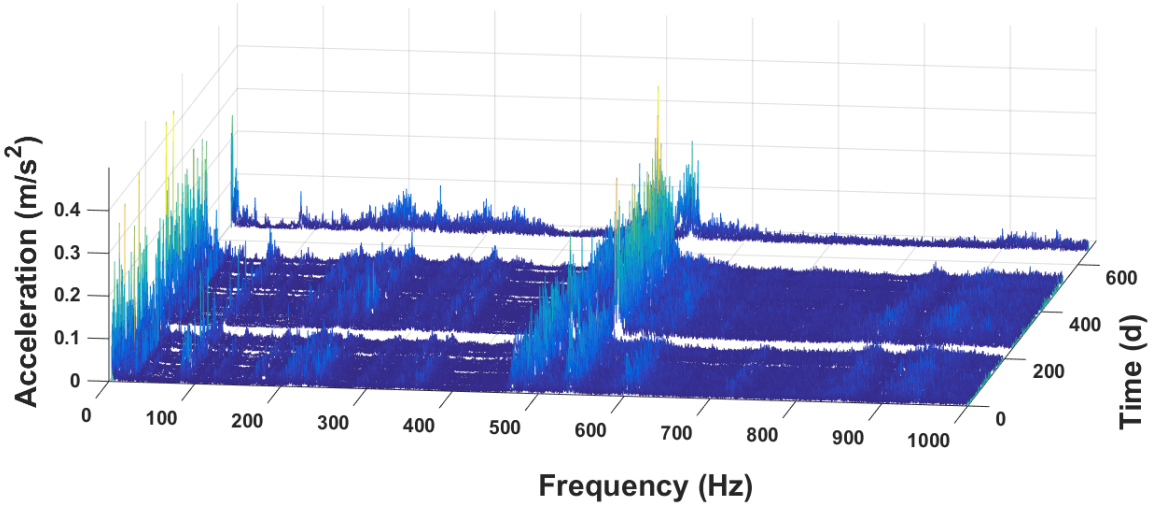


Figure 3. Waterfall plot of the spectra from the point RV in the frequency range 3 - 1000 Hz

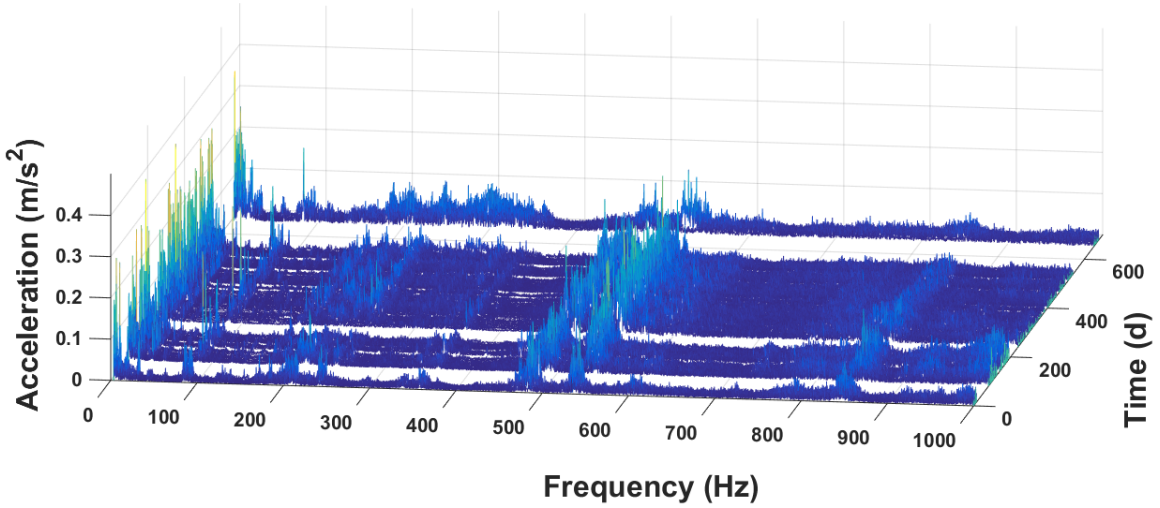


Figure 4. Waterfall plot of the spectra from the point LV in the frequency range 3 - 1000 Hz

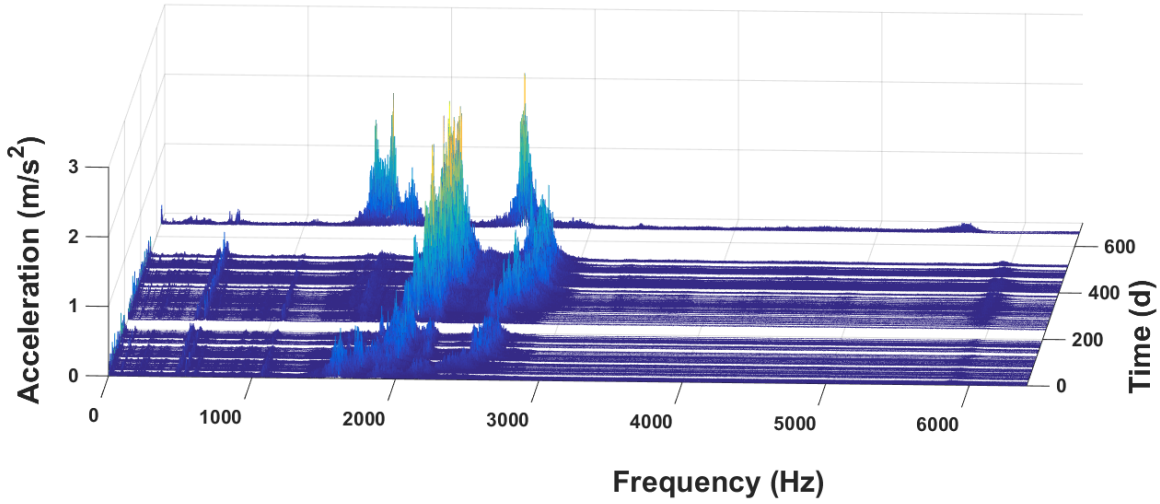


Figure 5. Waterfall plot of the spectra from the point RV in the frequency range 3 - 6400 Hz

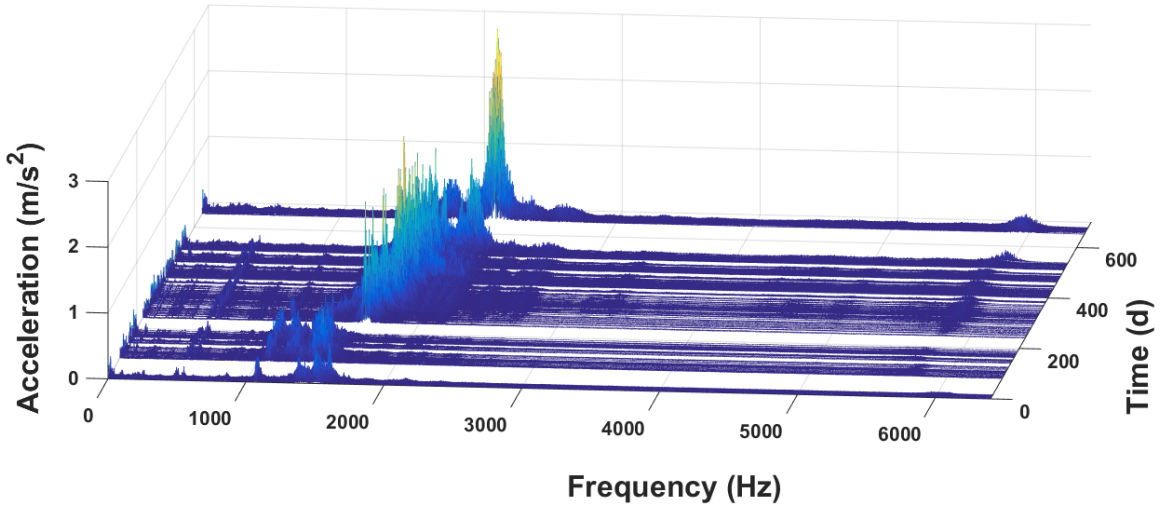


Figure 6. Waterfall plot of the spectra from the point LV in the frequency range 3 - 6400 Hz

Figure 2 shows the trend of values ${}^{2.4}\boldsymbol{MIT}_4^2$ from the point LV and fitted regression estimates with $h = 20$ and $h = 2$. After 350 days the trend starts to decrease but then makes a huge increase during the last gap in the measurements (or just before it, since there is one measurement before the gap for which ${}^{2.4}\boldsymbol{MIT}_4^2$ is almost 7). The trend estimate with $h = 2$ is shown here because it manages to estimate this last big change before the gap based on that one measurement.

Fig. 3, 4, 5 and 6 show waterfall plots of the spectra which were used in the calculation of the snap signals for the trends in the previous figures. In the frequency range 3 - 1000 Hz hardly any changes occur during the measurement period. This is interesting, since this frequency range contains the mesh frequencies (42.4 Hz and 121.5 Hz), their multiples and other gear related vibrations. In the higher frequencies we see huge increase in vibration amplitude around 2000 Hz and also a drift towards higher frequencies. These spikes are probably structural resonances and one can actually see with more in depth analysis that they are mostly very high order multiples of the cardan axle frequency 13.5 Hz. There is also increase in the 6000 Hz region with time. At the very end of the measurement period these high frequency resonances drop to smaller frequencies at the point RV (Fig. 5).

Some information on the damage in the axle after its breakdown was delivered to the university. It seems that the planetary gears had only minor wear on their surfaces. The ring gear of the differential had several pieces broken off from its teeth (probably explains the increase in the high multiples of the cardan frequency). There was also a lot of wear around the shafts in the axle and at least one cracked inner ring of a bearing.

5 Calculations from the WPS measurements

An overview of these measurements is provided in Fig. 7 where we have calculated the rms values from all of the vibration measurements together with the power data. It is clear that the WPS is operated very differently at different times. In the middle of summer the WPS is shut down during nights. The power output and vibration power seem to correlate so much that it is useful to investigate their relation in more detail.

Fig. 8 shows that the relation between acceleration rms and WPS power output is almost linear in gearbox 1. A good regression fit is achieved with parameter $h = 0.2$. Similar linear relationships between certain vibration frequency components and WPS power output were found in a previous study from the same WPS (Nikula et al, 2015). There is a tiny flatter part around 5 MW. This flat part is more defined in the similar visualisation from gearbox 2 in Fig. 9. It seems that the higher speed gearbox 2 (which also has much bigger vibration values than gearbox 1) also exhibits more nonlinearity in WPS power vs vibration power. These figures could be used to normalise vibration measurements taken during different power levels of WPS in its condition monitoring. One must be careful with such methods though, since for example if the measurements from low power situations are amplified, we will also decrease the signal to noise ratio in those cases when compared to those signals which are not amplified. In what follows we will consider only measurements from the power band 7.9 - 8.1 MW, since then we don't need such normalisation and the four continuous measurement periods all contain measurements from this power band.

From each signal from the power band 7.9 - 8.1 MW we took a 6.5536 second sample (to get 6.5536 s * 5000 Hz = 32768 samples, a power of two) and multiplied it with the window function (5) using $\varepsilon = 10$ and this new signal was differintegrated with the algorithm (4). High-pass filtering was performed with an ideal filter at cut-off frequency 3 Hz. Because the sampling frequency was relatively low, no low-pass filtering was done. From each end of the signal 20% was rejected and the remaining 3.9 second signal was used in the calculation of generalised l_p-norms.

Fig. 10 shows the trend of values ${}^{3.9}\boldsymbol{MIT}_2^8$ from gearbox 1 and a fitted regression estimate with $h = 20$. We see a 10 % increase in the last measurement period. Here and in the following trends we have a used a higher order norm just to show their effectiveness. In these calculations the norms with $p = 8$ showed a slightly more clear increase when compared to $p = 2$, but as we see in Fig. 11, the order of derivative plays a bigger role since the increase of ${}^{3.9}\boldsymbol{MIT}_4^8$ is 20 % and there is no downturn in July as there is in Fig. 10. The trend regression estimates with $h = 20$ and $h = 10$ only differ between the last two measurement periods as the smaller h shows a more rapid increase.

Same trends are calculated from the measurements from gearbox 2 and plotted in Fig. 12 and 13. Both show a 15 % increase during the measurement period, but the changes in the norm ${}^{3.9}\boldsymbol{MIT}_2^8$ are more irregular. This is especially clear in the trend regression estimate with $h = 10$ in Fig. 12. The bigger h ignores these local changes and reveals the long term trend more clearly. Even so, again the snap signals are more consistent as both trend regression estimates in Fig. 13 reveal the increasing trend better than the calculations from the acceleration signals.

Waterfall plot from gearbox 1 (Fig. 14) has increased at the very end of the measurement period at 1350, 1800 and 2250 Hz, which are multiples of $\nu_{meshWPS2}$. They also have sidebands 18 Hz apart, which is $\nu_{planetsWPS2}$. The gearbox 1 mesh frequency 117.3 Hz is also visible and it has sidebands 1.45 Hz apart (not visible in this zoom level of the figure), which is $\nu_{carrierWPS1}$. These do not change noticeably in the measurement period.

Waterfall plot from gearbox 2 (Fig. 15) shows hardly any change at all with time. It is a mystery why the increase at the frequencies 1800 and 2250 Hz are not visible here.

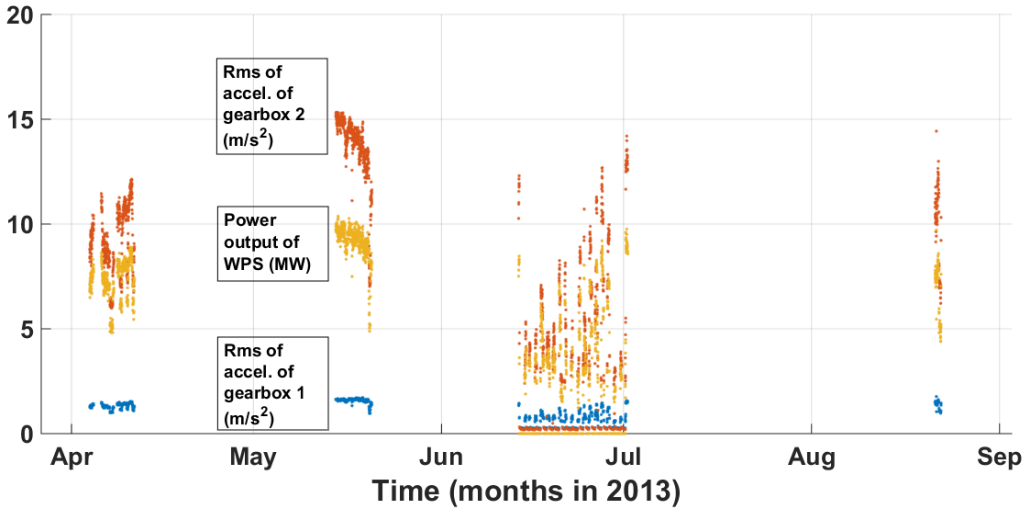


Figure 7. Overview of the measurements as rms of acceleration signals from both gearboxes and power output of WPS as functions of time

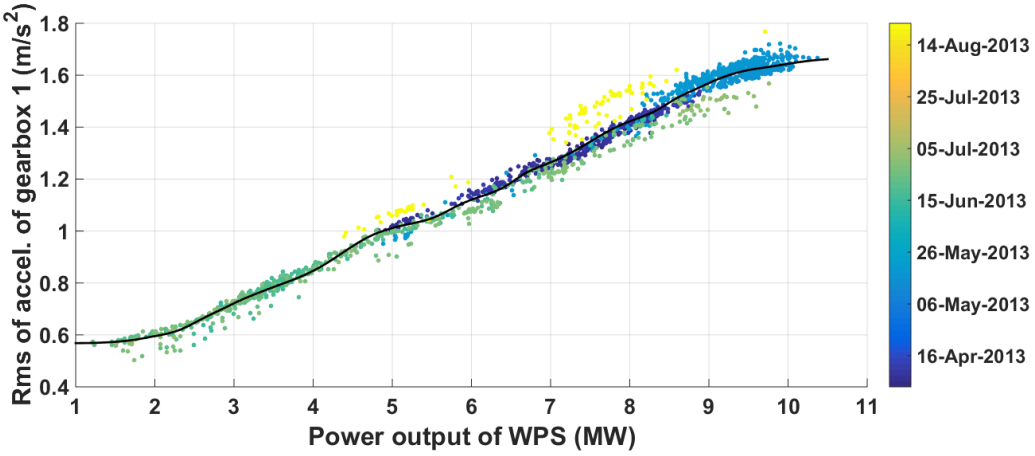


Figure 8. Power vs rms of acceleration from gearbox 1 and regression estimate with $h = 0.2$, colors depict time of measurements

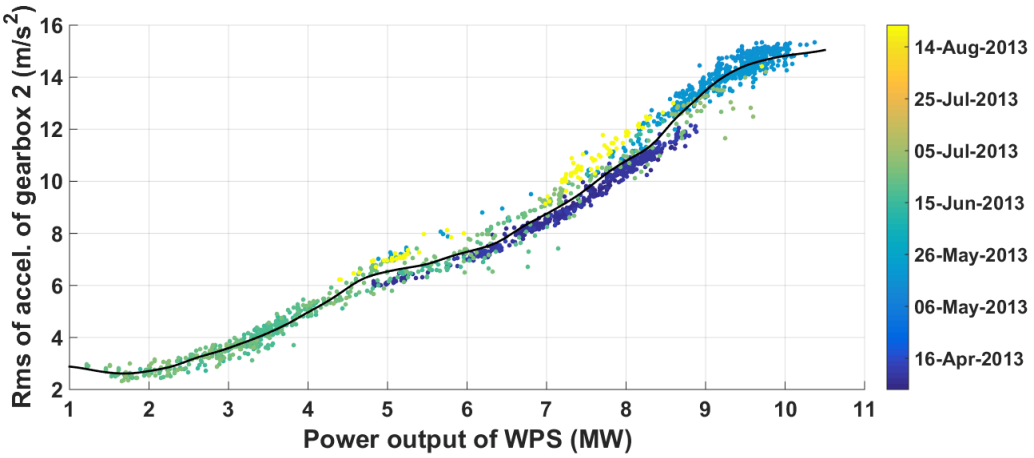


Figure 9. Power vs rms of acceleration from gearbox 2 and regression estimate with $h = 0.2$, colors depict time of measurements

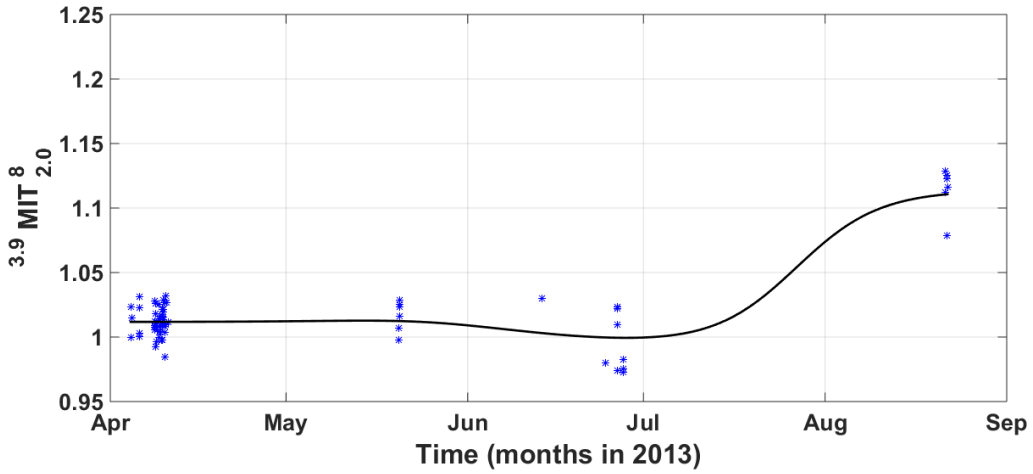


Figure 10. Trend of ${}^{3.9}\boldsymbol{MIT}_2^8$ from gearbox 1 in the frequency range 3 - 2500 Hz and regression estimate with $h = 20$

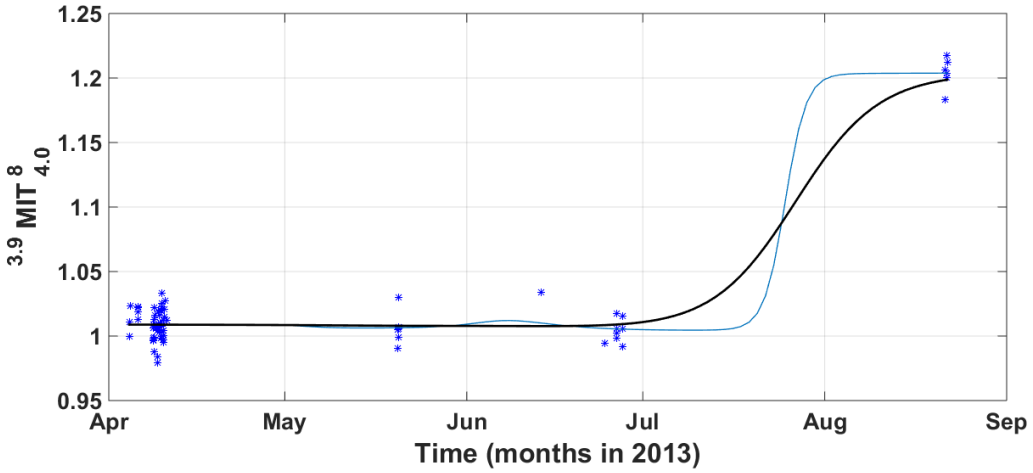


Figure 11. Trend of ${}^{3.9}\boldsymbol{MIT}_4^8$ from gearbox 1 in the frequency range 3 - 2500 Hz and regression estimates with $h = 20$ (black) and $h = 10$ (blue)

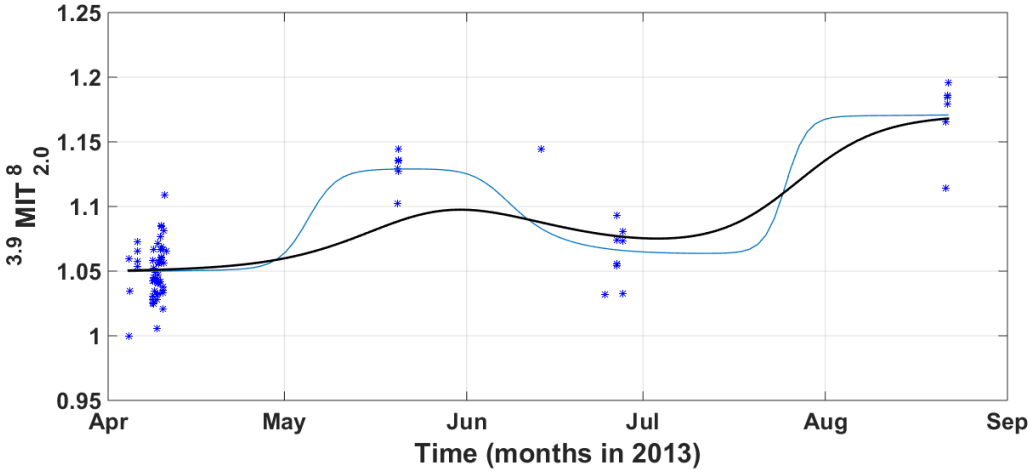


Figure 12. Trend of ${}^{3.9}\boldsymbol{MIT}_2^8$ from gearbox 2 in the frequency range 3 - 2500 Hz and regression estimates with $h = 20$ (black) and $h = 10$ (blue)

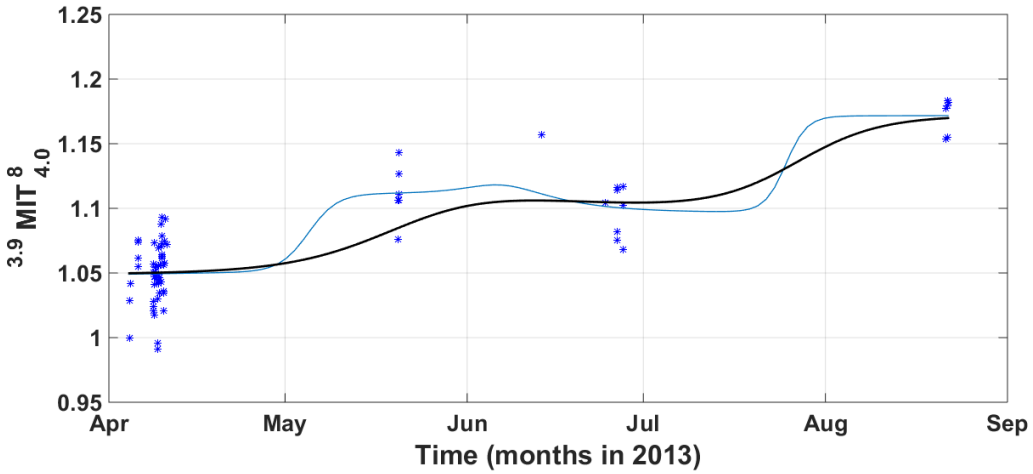


Figure 13. Trend of ${}^{3.9}\boldsymbol{MIT}_4^8$ from gearbox 2 in the frequency range 3 - 2500 Hz and regression estimates with $h = 20$ (black) and $h = 10$ (blue)

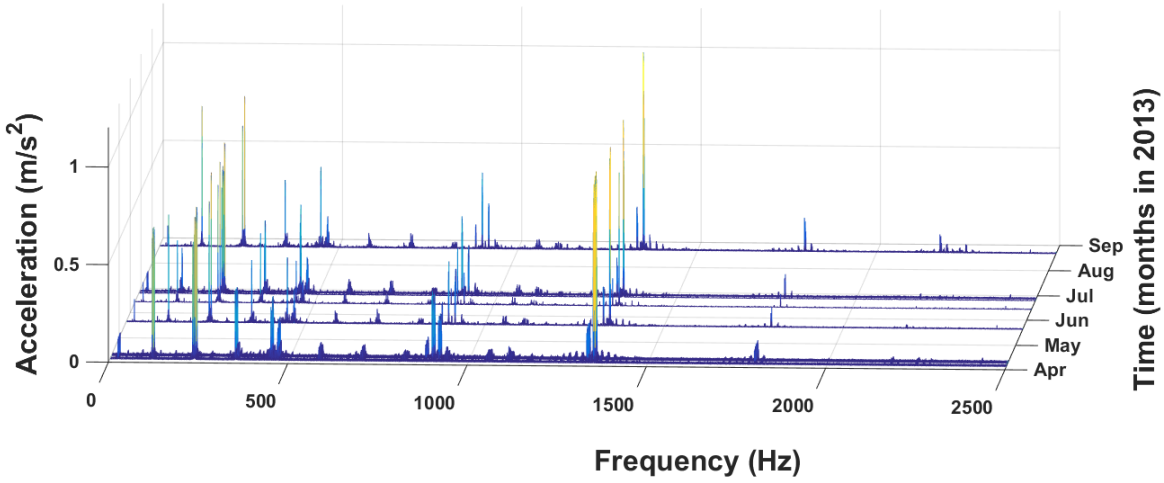


Figure 14. Waterfall plot of the spectra from gearbox 1 in the frequency range 3 - 2500 Hz

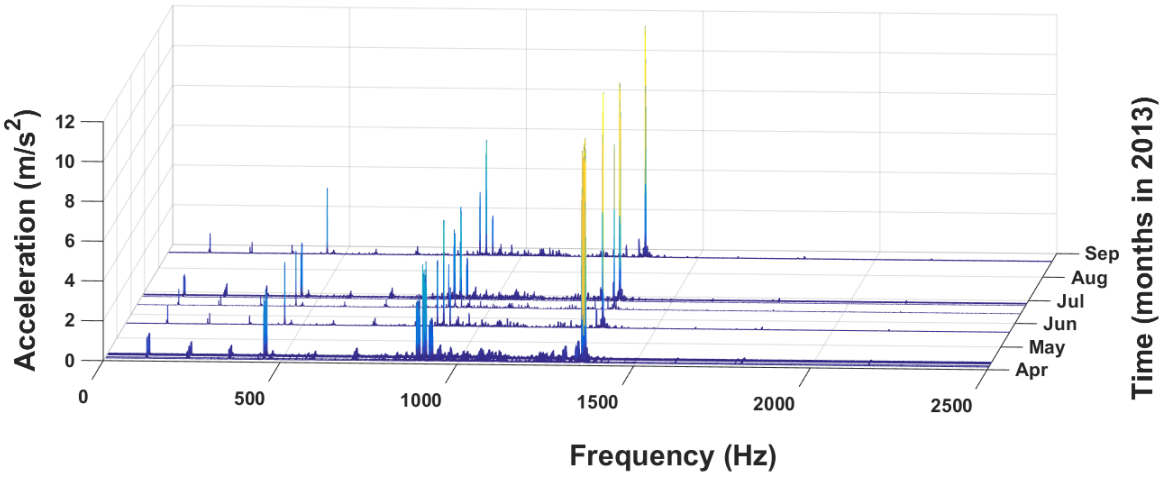


Figure 15. Waterfall plot of the spectra from gearbox 2 in the frequency range 3 - 2500 Hz

6 Conclusion

The long measurement period and the final breakdown of the front axle of the LHD makes it quite clear that vibration severity calculations could serve as indicator's of its condition. Snap signals are especially sensitive for deteriorating condition of the axle. Most of the critical faults occurred in the differential gearbox, which might explain why the spectral analysis only shows noticeable increase in the multiples of the cardan frequency.

Unfortunately the data from the WPS was intermittent and lasted only for one summer period, so we can not say for certain how big are for example its typical seasonal changes in vibration severity. An increasing trend was observed from both gearboxes especially when *MIT*-indices were calculated from snap signals. An increase in the multiples of the gearbox 2 mesh frequency was also observed, but interestingly from the gearbox 1 vibration measurements.

Acknowledgement

The LHD measurements were part of the "Development of production integrated condition-based maintenance model for mining industry (DEVICO)" project. The measurements at the WPS were part of the "Integrated condition-based control and maintenance (ICBCOM)" project. The first author wishes to thank Tauno Tönning foundation for their support on his doctoral studies.

References

W. Briggs and V. E. Henson. *The DFT An Owner's Manual for the Discrete Fourier Transform*. Society for Industrial and Applied Mathematics, 1995. ISBN 978-0898713428.

P. S. Bullen. *Handbook of Means and Their Inequalities*, 2nd ed. Kluwer Academic Publishers, 2003. ISBN 978-1402015229.

J. Immonen, S. Lahdelma and E. Juuso. Condition monitoring of an epicyclic gearbox at a water power station. In *Proc. The 53rd Scandinavian Conference on Simulation and Modelling (SIMS2012)*, Reykjavik, Iceland, Oct. 2012, pp. 99–105.

S. Lahdelma. *New vibration severity evaluation criteria for condition monitoring*. In Finnish, University of Oulu, Finland: Research report No 85, 1992, 18 pp.

S. Lahdelma and V. Kotila. Complex Derivative - A New Signal Processing Method. *Kunnossapito*, 19(4):39–46, 2005.

S. Lahdelma, and E. Juuso. Signal processing and feature extraction by using real order derivatives and generalised norms. Part 1: Methodology. *The International Journal of Condition Monitoring*, 1(2):46–53, 2011.

A. Laukka, J. Saari, J. Ruuska, E. Juuso and S. Lahdelma. Condition based monitoring for underground mobile machines. *International Journal of Industrial and Systems Engineering*, 23(1):74–79, 2016.

E. A. Nadaraya. On Estimating Regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964.

R.-P. Nikula, K. Leiviskä, K. Karioja. Epicyclic gearbox monitoring in a hydroelectric power plant with varying load. In *Proc. 11th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies (CM 2015 and MFPT 2015)*, Oxford, UK, Jun. 2015, 12 pp.

J. Nissilä, S. Lahdelma, and J. Laurila. Condition monitoring of the front axle of a load haul dumper with real order derivatives and generalised norms. In *Proc. 11th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies (CM 2014 and MFPT 2014)*, Manchester, UK, Jun. 2014, 20 pp.

J. K. Strayer. *Elementary Number Theory*. Waveland Press, 1994, Sect. 2.2.

C. M. Vicuña. *Contributions to the analysis of vibrations and acoustic emissions for the condition monitoring of epicyclic gearboxes*. Aachener Schriften zur Rohstoff- und Entsorgungstechnik des Instituts für Maschinentechnik der Rohstoffindustrie, Verlag R. Zillekens, Aachen, 2010.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

The Effect of Steel Leveler Parameters on Vibration Features

Riku-Pekka Nikula[1] Konsta Karioja[2]

[1]Control Engineering, University of Oulu, Finland, riku-pekka.nikula@oulu.fi
[2]Mechatronics and Machine Diagnostics, University of Oulu, Finland, konsta.karioja@oulu.fi

Abstract

The development of steel products with various characteristics increases the need for timely and preventive maintenance and condition monitoring of the production machinery. For instance, the roller levelers at modern steel factories are exposed to a high variation of forces due to the large range of steels leveled. In this study, the vibration measured from a steel leveler used for cold steel strips is analyzed with the goal to identify the effects different operational conditions have. Features such as generalized norms, generalized norm sums and the crest factor are computed from the vibration signals. The effects of the steel strip properties and the operational parameters of the machine on these features are then analyzed. The obtained information can be utilized in the models that are used as planning tools for the preventive maintenance of the steel leveler.

Keywords: feature extraction, roller leveling, signal processing, vibration

1 Introduction

The machines and devices in the modern steel industry are exposed to a high variation of load. This is a common characteristic especially at the factories which produce special steels. The special steel strips have properties such as exceptionally high yield strength. Major stresses are therefore inflicted on the processing equipment such as roller levelers during the production. This increases the risks of damage. The major forces make the roller leveler behave in an undesirable manner which makes, e.g., the mechanical load limiters break and the work rollers slip. Other common detriments encountered in the steel levelers include bearing damage in the rollers and the breakage and abrasion of the work rollers. These factors may consequently weaken the quality of the final steel product or cause a notable production loss due to the maintenance time. Therefore, the real-time monitoring of machine condition and the prediction of the effects of a specific steel product on the leveler are important subjects of research.

In this paper, the effects of the operating conditions during the steel leveling are studied from the vibration signal measured from the bodywork of a steel leveler. The studied roller leveler is used for strips of cold steel in a sheet line. The sheets are cut from the strip next to the leveler using a flying shear. The cutting causes impacts which are conducted to the leveler and the measured signal. Figure 1 depicts two examples of the measured signal during the processing of steel strips in the steel leveler. Figure 1 indicates that the effects of the steel strip properties and the operational parameters of the leveler are divergent in these cases. Differences between the leveling events can be commonly seen in the general signal level, in the impact magnitude of the sheet cutting and in the duration of the leveling events.

The identification of changes in the behavior of the monitored system can be done by using features (Lahdelma and Juuso, 2011). In this study, derivatives are first calculated in order to magnify the effects appearing in the signal. This is done according to the definitions by (Lahdelma, 1997) for real order derivatives. The actual vibration features are then extracted from the signal using the generalized norms introduced by (Lahdelma and Juuso, 2008). These norms are also used as the basis for other features such as generalized norm sums and the crest factor.

The signal derivatives and generalized norms have been previously used to demonstrate that different steel grades inflict different stress levels on a steel leveler (Karioja *et al*, 2015). Features based on the generalized norms have been proposed also as stress indicators for the same application (Nikula *et al*, 2017). In contrast to these studies, this paper addresses the effects of the operational parameters of the machine on the measured vibration. The effects of the steel strip properties on the general signal level and on the relative peak magnitude are studied as well. The correlations of the signal features with the machine parameters and steel strip properties are studied using Pearson's correlation coefficient. This information is useful for the development of data-driven models that are used in the preventive maintenance of the roller levelers.

2 Materials and Methods

2.1 Steel Leveler

The purpose of roller leveling is the elimination of various shape defects in the material. Steel coils contain flatness defects caused by uneven stresses and defects resulting from thickness variation across the product width (Smith, 1997). The stress patterns create

transverse and longitudinal curvature. Center and edge waves are caused by difference in the length of sheet between the center and the edges (Park and Hwang, 2002). Roller leveling is done by subjecting the strip to multiple back and forth bending sequences with decreasing roll penetrations as illustrated in Figure 2. The principle of roller leveling is based on controlling the plastic deformation through the thickness of the material. The plastic deformation determines the resultant flatness and memory and it also affects the required force. The roll force is a function of material thickness, width, yield strength, roll spacing and the extent of plastic deformation (Smith, 1997). A proper combination of the operational parameters is therefore needed for the required leveling result.

A schematic representation of a roller leveler is shown in Figure 2. The leveler under investigation is a four-high leveling machine used for strips of cold steel in a sheet line. The sheet cutting is performed simultaneously with the leveling without a need to stop the strip in the leveler due to the cutting. The cutting causes impacts that are conducted to the leveler and emerge as peaks in the monitored vibration signal which is also demonstrated in the lower graph in Figure 1.

2.2 Steel Leveler Parameters

The studied operational parameters of the leveler are presented in Table 1. The leveling of one complete steel strip is considered here as a leveling event. Exact time stamps for the operational parameters were unavailable and therefore the medians of each parameter represent the whole leveling event. The median represents the general level of an operational parameter during the leveling event. To precisely identify the instantaneous effect of an operational parameter on the vibration signal, the exact synchronization of time stamps is required.

The studied steel strip properties include yield strength, strip length, strip weight, strip width and strip thickness. The properties of the studied steel strips varied extensively. The range of the yield strength was approximately 200–1600 MPa; the length range was 68–1161 m; the thickness range was 1.98–15.21 mm; and the number of cut sheets was 4–465. The vibration signals from the leveling of altogether 739 steel strips were analyzed. The most common steel grade from 53 steel grades was a cold formable steel grade with yield strength around 400 MPa. This steel grade was leveled 123 times.

2.3 Vibration Measurements

The measurements were done at the SSAB rolling mill in Raahe, Finland. The accelerometer was stud-mounted in the middle of the runway on the bodywork supporting the lower supporting rolls. The acceleration was measured horizontally in the cross direction relative to the direction of the roller track. The used accelerometer was SKF CMSS 787A-M8, which has the frequency response from 0.7 Hz to 10 kHz with ±3 dB deviation. The measurement hardware included NI 9234 data acquisition card and NI CompactRIO for the data recording. The sampling rate was 25.6 kHz and the only filter used in hardware level was the built-in antialiasing filter of the data acquisition card. The measurement system was calibrated after the completion of measurements using a hand-held calibrator.

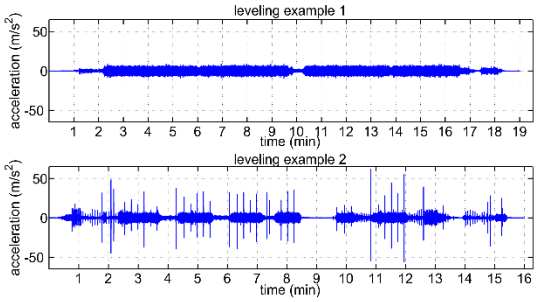


Figure 1. The acceleration signals from the leveling of two steel strips.

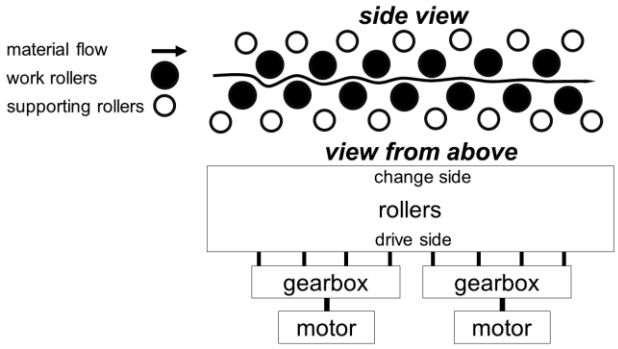


Figure 2. A schematic view of a roller leveler.

Table 1. The Studied Operational Parameters.

operational parameter	*parameter identifier*	*measure*
exit gap, drive side	*P*1	mm
exit gap, difference between sides	*P*2	mm
exit gap, change side	*P*3	mm
degree of plastic deformation	*P*4	%
entrance gap, drive side	*P*5	mm
entrance gap, difference between sides	*P*6	mm
entrance gap, change side	*P*7	mm
drawing force caused by roller leveler	*P*8	kN
drawing force caused by entry pinch roller	*P*9	kN

2.4 Signal Processing

The signals were processed using computational methods which are integration, derivation and filtering. The processing was done in the frequency domain by manipulating the sequence of complex numbers resulting from fast Fourier transform (FFT). Lahdelma (Lahdelma, 1997) defined real order derivative $x^{(\alpha)}$ of the function $x(t) = Xe^{i\omega t}$ in the form

$$x^{(\alpha)} = \omega^{\alpha} X e^{i\left(\omega t + \alpha\frac{\pi}{2}\right)}, \quad (1)$$

where α is the order of derivative, ω is angular frequency, X is amplitude, e is the Napierian number, i is the imaginary unit and t is the time variable. The derivative with respect to time can be calculated by multiplying every term X_k of the FFT by $(i\omega_k)^{\alpha}$ and then using the inverse of FFT. The integrals can be calculated similarly by using negative values of α. Furthermore, the value is not limited to integers but any real or complex number can be used (Lahdelma, 1997; Lahdelma and Kotila, 2005). Signals are filtered by multiplying the unwanted frequency components by zero.

In this study, the order of derivation was $\alpha = \{1, 2, 3, 4\}$. The variable $x^{(2)}$ corresponds to acceleration and x stands for displacement. The velocity signal ($\alpha = 1$) was filtered so that it included the frequencies 2–1000 Hz. The other signals were filtered only by the antialiasing filter embedded in the data acquisition card. The calculation of derivatives was done only for the leveling events that took less than 60 minutes. The events with longer duration were removed. Otherwise, the pre-processing was done according to the procedure presented in (Nikula *et al*, 2017). Five per cent of data was removed from the start and the end of each event to remove the effect of windowing in signal processing.

2.5 Vibration Features

The generalized norm is defined by

$$\left\|X^{(\alpha)}\right\|_p = \left[\frac{1}{N}\sum_{i=1}^{N}\left|x_i^{(\alpha)}\right|^p\right]^{\frac{1}{p}}. \quad (2)$$

This feature is known as the l_p norm of signal $x^{(\alpha)}$ where p is the order of the norm, α is the order of derivation and N is the number of signal values. The l_p norm has the same form as the generalized mean which is also known as the Hölder mean or power mean (Bullen, 2003). The root mean square and the peak value are special cases of the norm (2) when $p = 2$ and $p = \infty$, respectively. The large order of the norm magnifies the effect of the peaks, whereas the small order of the norm diminishes them.

The rms (l_2) was used to study the effect of the machine parameters and strip properties on the general signal level in this study. The crest factor (C) is the ratio between the absolute peak value (l_∞) and rms (l_2). The crest factor was used to study the relative magnitude of the impacts seen in the signal. These two features were computed from the segments, which had the duration of ten seconds. The mean of segments from the complete event was then used in the analysis of the effects.

The sums of $l_{0.1}$, l_2, l_4, l_{10}, and $l_1 + l_{10}$ were used to study the effects of the operational parameters of the machine only. A study concerning the correlations between the steel strip properties and the generalized norm sums is presented in (Nikula *et al*, 2017). The generalized norms were calculated using one-second samples in this case. Each sum was divided by the mean of the corresponding sums from the events, during which the most common steel grade was processed. Therefore, the value 1 corresponds to the mean of the most common steel grade in each sum. This operation was done in order to make the summation of l_1 and l_{10} practical.

2.6 Pearson's Correlation Coefficient

The correlation coefficient, R_{xy}, is defined by

$$R_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \quad (3)$$

where x_i and y_i are two different variables; $\bar{x}$ and $\bar{y}$ are their corresponding sample means; and n is the total number of observations. R_{xy} close to ±1 indicates strong linear correlation whereas R_{xy} close to 0 indicates weak linear correlation between the variables.

3 Results and Discussion

3.1 Correlations of Operational Conditions

Table 2 shows the correlation coefficients between the operational parameters of the machine and the steel strip properties based on the 739 events studied. The yield strength had a rather strong negative correlation with the gap values ($P1$, $P3$, $P5$, and $P7$), a strong negative correlation with the plastic deformation ($P4$) and a strong positive correlation with the drawing force caused by the machine ($P8$). The strip length had rather similar relationships with these operational parameters. The negative correlations with gap values were even stronger, but the correlations with the plastic deformation and the drawing force caused by the machine were slightly weaker. The correlations of strip weight and width with the operational parameters were generally weaker. The strong correlation of the strip width with the drawing force caused by the entry pinch roller ($P9$) is an exception. The correlations of the strip thickness were the opposite to the correlations of the yield strength or the length with the same operational parameters. This means that the thickness has a strong positive correlation with the gap values and the plastic deformation but quite strong negative correlation with the drawing force caused by the machine.

Table 2 roughly illustrates how the properties of the steel strips influence the operation of the roller leveler. In practice, the operational parameters are set based on the steel properties and then manipulated during the

leveling by the operator in the steelworks. Moreover, the strong correlations in Table 2 indicate that the medians of the operational parameters represent the operation sufficiently for the analysis purpose.

3.2 Correlations of Steel Strip Properties with Vibration Features

Table 3 shows the correlation coefficients of rms and crest factor with the steel strip properties. The rms calculated from the acceleration signal $x^{(2)}$ correlated the strongest with the yield strength, whereas the other steel strip properties had weak correlation. The rms values from the higher order derivatives of the signal indicated similar behavior. In the case of the velocity signal $x^{(1)}$, the correlation with the yield strength was slightly lower. The results indicate that the general signal level correlates the strongest with the yield strength considering the studied steel strip properties.

Figure 3 shows the rms of the $x^{(2)}$ signal together with two steel strip properties that had the strongest correlation. This 3D scatter plot illustrates that the strips with high yield strength are relatively thin. The thickness seems to have a positive correlation with rms if the yield strength is fixed although the general correlation is negative as shown in Table 3.

The crest factor had negative correlation with the strip length and positive correlation with the thickness according to Table 3. This result indicates that a thick steel strip results in relatively large impacts appearing in the signal. These impacts are mostly resulting from the sheet cutting using the flying shear. The weight and the width of the strip had low correlations in general. The effect of the yield strength varied based on the order of derivation. When the strip is thick, the yield strength and the strip length are relatively low, which explains the negative correlations of these two variables with the crest factor. When the velocity signal was used, the effect of weak impacts reduced on the crest factor. On the other hand, the strong impacts clearly stood out. This effect is shown in Figure 4. When the other signal derivatives were used, also the weaker impacts were magnified. This behavior is illustrated in the lower part of Figure 4. The level of the crest factor was lower in the case of $x^{(1)}$ signal compared with the higher order derivatives of the signal.

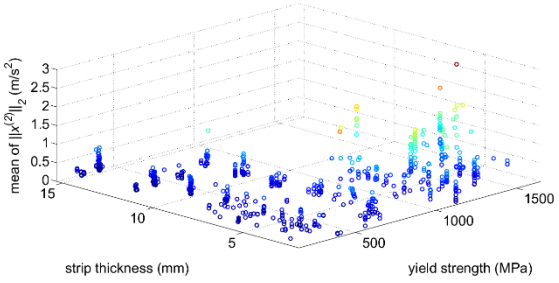


Figure 3. The effect of yield strength and thickness on the rms calculated from the acceleration signals.

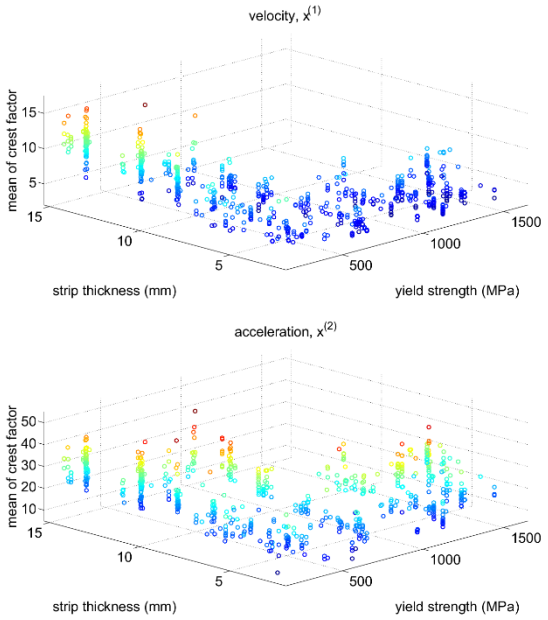


Figure 4. The effect of yield strength and thickness on the crest factor using $x^{(1)}$ and $x^{(2)}$ signals.

Table 2. Correlations between the Operational Parameters and the Steel Strip Properties.

id.	*yield strength*	*length*	*weight*	*width*	*thickness*
P1	-0.656	-0.793	0.441	0.277	0.991
P2	0.218	-0.075	0.070	0.093	-0.017
P3	-0.662	-0.790	0.438	0.274	0.990
P4	-0.875	-0.630	0.401	0.286	0.758
P5	-0.673	-0.810	0.490	0.332	0.973
P6	-0.036	-0.209	0.151	0.159	0.153
P7	-0.676	-0.809	0.488	0.329	0.973
P8	0.853	0.510	-0.312	-0.183	-0.683
P9	-0.116	-0.201	0.561	0.931	0.156

Table 3. Correlations of the Steel Strip Properties with rms and Crest Factor.

feature	*yield strength*	*length*	*weight*	*width*	*thick-ness*
$\|\|x^{(1)}\|\|_2$	0.487	0.048	0.094	0.039	-0.085
$\|\|x^{(2)}\|\|_2$	0.660	0.192	0.072	0.056	-0.310
$\|\|x^{(3)}\|\|_2$	0.630	0.207	0.031	0.021	-0.328
$\|\|x^{(4)}\|\|_2$	0.634	0.217	0.022	0.013	-0.339
C, $x^{(1)}$	-0.516	-0.569	0.275	0.175	0.759
C, $x^{(2)}$	0.091	-0.401	0.248	0.099	0.384
C, $x^{(3)}$	-0.217	-0.422	0.179	0.126	0.470
C, $x^{(4)}$	-0.251	-0.423	0.185	0.154	0.468

3.3 Correlations of Operational Parameters with Vibration Features

Table 4 shows the correlation coefficients of rms and crest factor with the operational parameters of the machine. The correlations of rms are shown using only $x^{(1)}$ and $x^{(2)}$ signals. The rms of $x^{(3)}$ and $x^{(4)}$ signals had almost the same correlations as the rms of $x^{(2)}$ signal. In general, the linear correlations between the operational parameters and rms were rather low. The drawing force caused by the leveler ($P8$) had the strongest correlation. As shown in Table 2, this parameter also had a strong correlation with the yield strength, which has a large influence on the rms level as shown in the previous Section.

The crest factor had a strong positive correlation with the roller gap values and the degree of plastic deformation when calculated from the $x^{(1)}$ signal. These correlations were weaker using the higher order derivatives of the signal. The drawing force caused by the roller leveler ($P8$) had a stronger negative correlation with the crest factor of $x^{(1)}$ signal compared with the crest factor of the other signals. The correlations between the yield strength and the crest factor of different signal derivatives indicated a similar behavior which is shown in Table 3.

Table 5 shows the correlations between the operational parameters of the machine and the generalized norm sums calculated from the $x^{(2)}$ signal. The generalized norm sums correlated the strongest with the drawing force caused by the roller leveler. The gap values and the degree of plastic deformation had substantially strong negative correlations. The results indicate that the correlations were the highest when the order of the norm was the lowest in these cases. In most cases, the second highest correlation was observed with the norm combination l_1+l_{10}.

The effects of operational parameters $P1$, $P4$, $P5$, and $P8$ on $l_{0.1}$ sum are depicted in Figure 5. According to the presented gap values, a thick steel strip causes a low accumulation of norm values. When the gap is reduced, the norm sums increase and an increasing variation in the sums is observed. This effect is seen in the graphs illustrating the effects of gap values and the degree of plastic deformation. The norm sums seem to increase together with the drawing force caused by the roller leveler. When the operational parameters are considered fixed, the variation in the norm sums can be explained by the effects of other variables. After all, the effects seen in the vibration are the result of the behavior of a multivariate system.

3.4 Discussion

The effects of the operational parameters of the steel leveler on the vibration signals were demonstrated in this study. Moreover, the results indicated that the operational parameters have strong correlations with the different steel strip properties. This implies that the characteristics of the vibration could be predicted with an approximate precision before the leveling is done based solely on the steel strip properties.

The results indicated that the steel strips with high yield strength inflict relatively high vibration level, which is seen in the rms values. According to the results in Table 4, the drawing force caused by the leveler apparently had an effect on this as well. Figure 3 also indicates that the steel thickness affects the rms value when the yield strength is kept fixed. This implies that the strip thickness could be used to predict the rms level during the processing of a specific steel grade because the yield strength is fixed in that case. The high level of rms may be an indication of the potential risks of slipping in the work rollers or abrasion, for instance.

The impacts caused by the sheet cutting manifested themselves in the crest factor especially with thick strips as shown in Figure 4. In addition, other variables such as the yield strength had an influence especially in the case of relatively thin strips. The roller gap values indicated a positive correlation with the crest factor as well. The impacts presumably stress the bearings of the work rollers and supporting rollers, which are under a major load during the leveling.

The roller gap values had quite strong negative correlation with the generalized norm sums. This indicates that the thin strips, which are also long, accumulate higher sums compared with the thick strips. This means that the leveler is stressed a relatively long duration. The generalized norm sums could be used for the general machine stress evaluation.

The roller gap differences between the sides ($P2$ and $P6$) had low correlations with the vibration features. These parameters had low correlations with the studied steel strip properties as well. The entry pinch roller ($P9$) had weak correlations with the vibration features, but a strong correlation with the strip width. The strip weight and the width had low correlations with the studied vibration features as well. These parameters apparently have poor applicability to the prediction of the studied effects appearing in the vibration signal. The effects of the positions of the supporting rollers were rejected in this study because information on the possible steel strip shape flaws was unavailable. In general, the supporting rollers are adjusted to correct flatness defects (Smith, 1997).

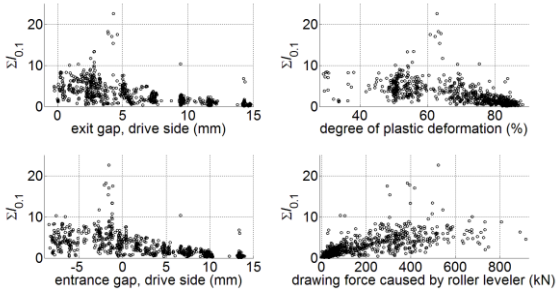


Figure 5. The effect of four operational parameters on $l_{0.1}$ sums using $x^{(2)}$ signals.

Table 4. Correlations of the Operational Parameters with rms and Crest Factor.

	$x^{(1)}$	$x^{(2)}$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
id.	*rms*		*Crest Factor*			
P1	-0.135	-0.366	0.760	0.344	0.457	0.459
P2	0.154	0.175	-0.075	0.222	0.008	-0.014
P3	-0.139	-0.371	0.762	0.338	0.456	0.459
P4	-0.280	-0.469	0.609	0.128	0.320	0.342
P5	-0.149	-0.350	0.734	0.365	0.467	0.470
P6	0.011	0.005	0.057	0.230	0.085	0.073
P7	-0.150	-0.352	0.736	0.361	0.467	0.471
P8	0.434	0.588	-0.522	-0.088	-0.289	-0.315
P9	0.059	0.087	0.115	0.059	0.072	0.091

Table 5. Correlations of the Operational Parameters with the Sums of the Generalized Norms.

id.	$\sum l_{0.1}$	$\sum l_2$	$\sum l_4$	$\sum l_{10}$	$\sum(l_1+l_{10})$
P1	-0.593	-0.511	-0.483	-0.484	-0.532
P2	0.108	0.141	0.160	0.168	0.145
P3	-0.595	-0.515	-0.487	-0.488	-0.536
P4	-0.622	-0.577	-0.561	-0.560	-0.593
P5	-0.600	-0.507	-0.473	-0.472	-0.528
P6	-0.080	-0.050	-0.026	-0.013	-0.044
P7	-0.601	-0.508	-0.475	-0.474	-0.529
P8	0.658	0.618	0.592	0.582	0.626
P9	0.002	0.051	0.061	0.057	0.041

4 Conclusions

The effects of steel leveler parameters on vibration features were studied based on the observed linear correlations. Signal derivatives were calculated to magnify the effects. Crest factor, rms and generalized norm sums were used as the vibration features. The crest factor, which shows the ratio of the peak amplitude and rms, had strong correlation especially with strip thickness and the roller gap parameters. The general magnitude of a signal, defined by rms, had the strongest correlations with the yield strength of strip and the drawing force caused by the roller leveler. The generalized norm sums, which can be used to indicate stress accumulation, had the strongest correlations with the drawing force caused by the roller leveler, the roller gap values and the degree of plastic deformation. The strong correlations between the steel strip properties and the operational parameters of the leveler suggest that the characteristics of the vibration signal could be predicted based on the strip properties solely. This implies that the steel strip properties could be used as input variables in models, which predict different effects on the vibration. These models could then be used in the preventive maintenance of the steel leveler.

Acknowledgements

The authors thank the personnel of SSAB Europe for collaboration and enabling of the measurement campaign during the SIMP (System Integrated Metals Processing) program coordinated by FIMECC (Finnish Metals and Engineering Competence Cluster).

References

P.S. Bullen. *Handbook of Means and Their Inequalities*, 2nd ed. Kluwer Academic Publishers. 2003.

Konsta Karioja, Riku-Pekka Nikula, and Toni Liedes. Vibration Measurements and Signal Processing in Stress Monitoring of a Steel Leveller. *Condition Monitoring and Diagnostics & Maintenance Performance Measurement and Management: MCMD 2015 and MPMM 2015*, Oulu, Finland, 2015.

Sulo Lahdelma. On the Derivative of Real Number Order and Its Application to Condition Monitoring. *Kunnossapito*, 17:39–42, 1997.

Sulo Lahdelma and Esko K. Juuso. Signal Processing in Vibration Analysis. *The Fifth International Conference on Condition Monitoring and Machinery Failure Prevention Technologies*, Edinburgh, p. 879-889, 2008.

Sulo Lahdelma and Esko Juuso. Signal Processing and Feature Extraction by using Real Order Derivatives and Generalised Norms. Part 2: Applications. *International Journal of Condition Monitoring*, 1(2):54–66, 2011. doi: 10.1784/204764211798303814.

Sulo Lahdelma and Vesa Kotila. Complex Derivative – A New Signal Processing Method. *Kunnossapito*, 19:39–46, 2005.

Riku-Pekka Nikula, Konsta Karioja, Kauko Leiviskä, and Esko Juuso. Prediction of Mechanical Stress in Roller Leveler Based on Vibration Measurements and Steel Strip Properties. *Journal of Intelligent Manufacturing*, 2017. doi: 10.1007/s10845-017-1341-3.

Kee-cheol Park and Sang-Moo Hwang. Development of a Finite Element Analysis Program for Roller Leveling and Application for Removing Blanking Bow Defects of Thin Steel Sheet. *ISIJ International*, 42(9):990–999, 2002. doi: 10.2355/isijinternational.42.990.

Richard P. Smith, Jr. Flatness Control in Coiled Plates: Lukens' Wide, Cut to Length Line. *Iron and Steel Engineer*, 74:29–34, 1997.

Spline Trajectory Planning for Path with Piecewise Linear Boundaries

Hiroyuki Kano[1] Hiroyuki Fujioka[2]

[1]Division of Science, Tokyo Denki University, Saitama 350-0394, Japan, kano@mail.dendai.ac.jp
[2]Department of System Management, Fukuoka Institute of Technology, Fukuoka 811-0295, Japan, fujioka@fit.ac.jp

Abstract

We consider a problem of trajectory planning for path with piecewise linear boundaries. The trajectory is constructed as smoothing splines using normalized uniform B-splines as the basis functions. The boundary constraints are treated as a collection of inequality pairs by right and left boundary lines, and are formulated as linear inequality constraints on the so-called control point vector. Smoothing splines are constructed as an approximation of a piecewise linear centerline of the given path, where the given entire time interval is divided into subintervals according to the centripetal distribution rule. Other constraints as initial and terminal conditions on the trajectory can be included easily, and the problem reduces to convex quadratic programming problem where very efficient numerical solvers are available. The effectiveness of the proposed method is confirmed by an example of fairly complex path with piecewise linear boundaries. Also an example is included to demonstrate its usefulness for trajectory planning in an environment with obstacles.

Keywords: trajectory planning, smoothing spline, B-spline, boundary constraint, quadratic programming problem

1 Introduction

Splines have been used frequently in robotics as in the problems of trajectory planning of robotic arms and mobile robots (Biagiotti and Melchiorri, 2008; Egerstedt and Martin, 2010; Khalil and Dombre, 2002). A typical problem of trajectory planning consists of constructing a function of time that satisfies initial and terminal conditions together with other requirements such as via points and obstacle avoidance.

When via points are specified, trajectories may be constructed as interpolating splines to pass the via points or as approximating splines to pass near the points (Crouch and Jackson, 1991; Egerstedt and Martin, 2001). The problems of obstacle avoidance trajectory planning are often treated by introducing a cost function consisting of distance to obstacles together with e.g. the trajectory length, which are expressed as nonlinear function of some parameters representing the trajectories. Cubic splines are frequently used to construct trajectories (Kolter and Ng, 2009; Saska et al., 2006; Piazzi and Visioli, 2000), and an optimization problem is solved numerically for trajectory planning. Particle swarm optimization method is employed in (Saska et al., 2006).

In (Gallina and Gasparetto, 2000), representing trajectories by sums of harmonics, the trajectory planning problem is formulated as constrained nonlinear programming problem, where the obstacles are treated as inequality constraints by assuming their parametric representation as polygons and ellipses. Also, treating obstacles as linear inequality constraints in (Berglund et al., 2010), nonlinear programming problem is solved for quartic B-spline curves with minimum curvature. Only B-splines of degrees two to four are allowed.

In this paper, we consider a problem of trajectory planning for road-like path with the right and left boundaries in a 2-dimensional plane. The boundaries are assumed to be given as piecewise linear functions, and the problem is to construct a trajectory from given start point to goal point without exceeding the boundaries. Although such a problem of planning trajectories for path with boundaries naturally arises, the treatment as in this study seems novel to the authors' knowledge.

To be more specific, we construct trajectories as smoothing splines using B-splines as the basis functions (Kano et al., 2005). This approach is very suitable to the present problem, since we can construct trajectories piecewise in accordance with each piece of boundaries. Also, by dividing the entire time interval into subintervals, the problem is to construct the trajectory bounded by two lines for each time interval.

In such problems, our approach by constrained splines (Kano et al., 2011; Fujioka and Kano, 2012; Kano et al., 2014) is very effective. The studies on constrained splines include constraints at isolated time instants, those over an interval of time, constraints on function values, on time derivatives of arbitrary degrees, or on integral values, and so forth. The constraints can be equality and/or inequality, and are systematically included in the formulation by B-spline based smoothing splines.

We show that the problem is formulated as a convex QP (quadratic programming) problem. The description of the problem is easy since the boundaries can be defined by simply providing a series of pairs of right and left cor-

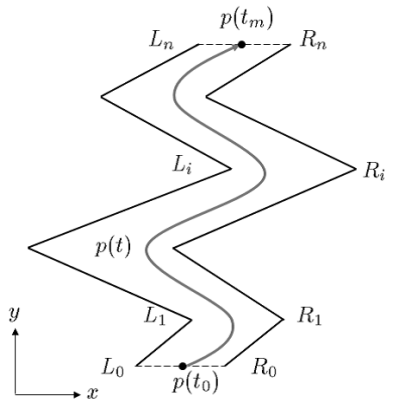


Figure 1. A path with piecewise linear boundaries.

ners, and the resulting QP problem is solved efficiently numerically by existing QP solver. The usefulness of the proposed method is confirmed by two examples: first, for relatively complex path with piecewise linear boundaries, and, second, for application to obstacle avoidance trajectory planning problem.

This paper is organized as follows. In Section 2, we present problem statement and describe 2-dimensional vector smoothing splines based on B-splines. Then in Section 3, the trajectory planning problem is formulated and solved. Two numerical exampled are considered in Section 4. Concluding remarks are given in Section 5.

Throughout the paper, the symbol $\otimes$ denotes the Kronecker product, and 'vec' the vec-function (see e.g. (Lancaster and Tismenetsky, 1985)).

2 Preliminaries

2.1 Problem Statement

As shown in Fig. 1, let a path with piecewise linear boundaries be defined by a pair of corner points (R_i, L_i), $i = 0, 1, \cdots, n$ on xy-plane. Then we consider to design a trajectory $p(t) \in \mathbf{R}^2$

$$p(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \tag{1}$$

for given time interval $[t_0, t_m]$ and for given initial position $p(t_0)$ and final position $p(t_m)$. Our particular interest is to construct a smooth trajectory $p(t)$ that is guaranteed to stay within the path specified by the piecewise linear right boundary $R_0R_1\cdots R_n$ and the left boundary $L_0L_1\cdots L_n$. It is noted that the initial and terminal conditions can be specified as equality and/or inequality conditions on $p(t)$ and its derivatives.

We construct the trajectory $p(t)$ for $t \in [t_0, t_m]$ by splines using normalized uniform B-spline $B_k(t)$ of degree $k(\geq 1)$,

$$p(t) = \sum_{i=-k}^{m-1} \tau_i B_k(\alpha(t - t_i)). \tag{2}$$

Table 1. $N_{j,3}(t)$ $(j = 0, 1, 2, 3)$ and its derivatives

j	$3!N_{j,3}(t)$	$2!N^{(1)}_{j,3}(t)$	$N^{(2)}_{j,3}(t)$	$N^{(3)}_{j,3}(t)$
0	$(1-t)^3$	$-(1-t)^2$	$1-t$	-1
1	$4-6t^2+3t^3$	$-4t+3t^2$	$-2+3t$	3
2	$1+3t+3t^2-3t^3$	$1+2t-3t^2$	$1-3t$	-3
3	t^3	t^2	t	1

Here $\tau_i \in \mathbf{R}^2$ are weighting coefficients called control points, and $\alpha(>0)$ is a constant for scaling the interval between equally-spaced knot points t_i with

$$t_{i+1} - t_i = \frac{1}{\alpha}. \tag{3}$$

Moreover, $B_k(t)$ is defined by

$$B_k(t) = \begin{cases} N_{k-j,k}(t-j) & j \leq t < j+1, \\ & j = 0, 1, \cdots, k \\ 0 & t < 0 \ \text{ or } \ t \geq k+1, \end{cases} \tag{4}$$

where the basis elements $N_{j,k}(t)$ $(j = 0, 1, \cdots, k)$, $0 \leq t \leq 1$ can be derived recursively by de Boor's algorithm (de Boor, 2001) for any $k \geq 1$. Note that this basis elements satisfies $N_{j,k}(t) \geq 0 \ \forall t \in [0, 1]$ and

$$\sum_{j=0}^{k} N_{j,k}(t) = 1, \ \ \forall t \in [0, 1]. \tag{5}$$

Since cubic splines are most frequently used, for reference, we show $N_{j,3}(t)$ together with its derivatives in Table 1. Smoothing splines are used to determine the control points τ_i, or the control point matrix $\tau \in \mathbf{R}^{2\times M}$ $(M = m + k)$

$$\tau = \begin{bmatrix} \tau_{-k} & \tau_{-k+1} & \cdots & \tau_{m-1} \end{bmatrix} \tag{6}$$

as we see in the sequel.

2.2 Vector Smoothing Splines

In this particular problem, we consider smoothing splines for continuous-time data $f(t) \in \mathbf{R}^2$, where the following cost function is minimized.

$$J(\tau) = \lambda \int_{t_0}^{t_m} \left\| p^{(l)}(t) \right\|_\Lambda^2 dt + \int_{t_0}^{t_m} \| p(t) - f(t) \|^2 dt. \tag{7}$$

Here $\lambda(>0)$ is a smoothing parameter, $\Lambda \in \mathbf{R}^{2\times 2}$ is a positive-definite weight matrix, $\|u\|^2 = u^T u$, and $\|u\|_\Lambda^2 = u^T \Lambda u$. We take the integer l as $l = 2$ for cubic spline $(k = 3)$ and $l = 3$ for quintic spline $(k = 5)$. Thus the problem is to construct a smooth spline $p(t)$ that approximate the function $f(t)$, which is given, typically so as to represent the center line of the path.

It is noted that usual smoothing spline problem (Wahba, 1990) employs discrete-time set of data (s_i, f_i) in which case the second term in (7) is set as $\sum_{i=1}^{N} \| p(s_i) - f_i \|_{W_i}^2$.

Now, let $\hat{\tau} \in \mathbf{R}^{2M}$ be the vec-function (Lancaster and Tismenetsky, 1985) of τ, i.e.

$$\hat{\tau} = \text{vec } \tau. \tag{8}$$

Then, following the similar procedure as in (Kano et al., 2005), $J(\tau)$ in (7) is expressed as a quadratic function $J(\hat{\tau})$ in $\hat{\tau}$,

$$J(\hat{\tau}) = \hat{\tau}^T G \hat{\tau} - 2\hat{\tau}^T g + g_c \tag{9}$$

where $G \in \mathbf{R}^{2M\times 2M}$, $g \in \mathbf{R}^{2M}$ and $g_c \in \mathbf{R}$ are given by

$$G = Q \otimes \Lambda + Q_0 \otimes I_2 \tag{10}$$

$$g = \int_{t_0}^{t_m} b(t) \otimes f(t) dt \tag{11}$$

$$g_c = \int_{t_0}^{t_m} \|f(t)\|^2 dt. \tag{12}$$

Here $b(t) \in \mathbf{R}^M$ is a vector of shifted B-splines defined by

$$b(t) = \left[\begin{array}{cc} B_k(\alpha(t-t_{-k})) & B_k(\alpha(t-t_{-k+1})) \\ \cdots & B_k(\alpha(t-t_{m-1})) \end{array} \right]^T, \tag{13}$$

and $Q, Q_0 \in \mathbf{R}^{M\times M}$ are Gram matrices defined by

$$Q = \int_{t_0}^{t_m} \frac{d^l b(t)}{dt^l} \frac{d^l b^T(t)}{dt^l} dt \tag{14}$$

$$Q_0 = \int_{t_0}^{t_m} b(t) b^T(t) dt \tag{15}$$

3 Trajectory Planning

Recalling that the path is constrained by piecewise linear boundaries as in Fig. 1, it is convenient to plan the trajectory $p(t)$ for each piece of the path constrained by a pair of right and left boundary line segments, R_iR_{i+1} and L_iL_{i+1}. The construction of $p(t)$ in (2) is very suitable for this purpose since it is a piecewise polynomial with the knot points t_i.

For this purpose, we divide the time interval $[t_0, t_m]$ into n subintervals $[s_i, s_{i+1}]$, $i = 0, 1, \cdots, n-1$ in accordance with n pairs of boundary segments R_iR_{i+1} and L_iL_{i+1}. Here we take $[s_i, s_{i+1}]$ to be a knot point interval, namely each s_i is taken as one of the knot point t_j with $s_0 = t_0$ and $s_n = t_m$.

Now denote the straight line passing two points R_i and R_{i+1} by l_i, and the one for points L_i and L_{i+1} by l'_i (see Fig. 2). Then we plan the trajectory $p(t)$ so that, for each $i = 0, 1, \cdots, n-1$, it lies between the two boundary lines l_i and l'_i for all t in $[s_i, s_{i+1}]$.

This condition is described in Section 3.1, a method of assigning s_i as a knot point in Section 3.2, and the spline construction procedure will be given in Section 3.3, e.g. taking initial and final conditions into account.

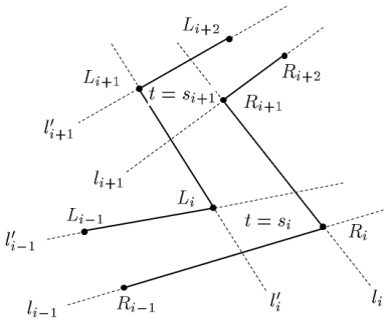


Figure 2. Corner points (R_i, L_i) and boundary lines (l_i, l'_i).

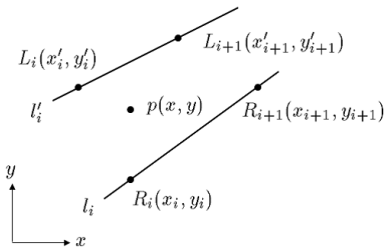


Figure 3. Right and left boundary lines l_i, l'_i and a point p.

3.1 Trajectory between Two Lines

For the present problem, it is natural to introduce the following assumptions (see Figs. 1 and 2).

(A1) The polygon $\mathscr{P}_i = R_iR_{i+1}L_{i+1}L_i$ is a convex quadrangle for all i.

(A2) In $\mathscr{P}_i$, the vertices $R_i, R_{i+1}, L_{i+1}, L_i$ are located counterclockwise.

By (A1), all the four points are distinct and the line segment R_iR_{i+1} does not intersect with L_iL_{i+1}. By (A2), the quadrangle $\mathscr{P}_i$ constitutes part of the path with piecewise linear boundaries, or the path is the union of the quadrangles.

Now we derive a condition such that $p(t)$ remains in a region between the two lines l_i and l'_i for all t in a knot point interval $[s_i, s_{i+1}]$, where we let $[s_i, s_{i+1}] = [t_\kappa, t_\mu]$with $\kappa < \mu$. Here and hereafter, by the term 'region between the two lines l_i and l'_i', we mean the region between the two lines including the quadrangle $\mathscr{P}_i$. Thus if lines l_i and l'_i intersect, then among the two wedge-shaped regions, the one including $\mathscr{P}_i$ is meant.

Since the lines l_i and l'_i respectively play the roles of right and left boundaries of the path, we require the point $p(t)$ to lie to the left of l_i when viewed from R_i toward R_{i+1} along l_i, and moreover, $p(t)$ to lie to the right of l'_i when viewed from L_i toward L_{i+1} along l'_i.

In view of Fig. 3, the above conditions for a point $p(x, y)$ to lie between the two lines can be written as

$$\left\{ \begin{array}{ll} \overrightarrow{R_ip} \times \overrightarrow{R_iR_{i+1}} & \leq 0 \\ \overrightarrow{L_ip} \times \overrightarrow{L_iL_{i+1}} & \geq 0. \end{array} \right. \tag{16}$$

Here $\times$ denotes cross product of 2-dimensional vectors, and it holds that, for $u = [u_1\ u_2]^T$ and $v = [v_1\ v_2]^T$

$$u \times v = \begin{vmatrix} u_1 & v_1 \\ u_2 & v_2 \end{vmatrix} = u_1 v_2 - u_2 v_1 \tag{17}$$

Applying this relation to the vectors in (16), for example $\overrightarrow{R_i p} = [x - x_i\ y - y_i]^T$, (16) can be rewritten as linear inequalities in $p = [x\ y]^T$ as follows.

$$Ap \le d, \tag{18}$$

where

$$A = \begin{bmatrix} -y_i + y_{i+1} & x_i - x_{i+1} \\ y'_i - y'_{i+1} & -x'_i + x'_{i+1} \end{bmatrix}, \tag{19}$$

$$d = \begin{bmatrix} x_i y_{i+1} - x_{i+1} y_i \\ -x'_i y'_{i+1} + x'_{i+1} y'_i \end{bmatrix}. \tag{20}$$

Next we consider the condition such that $p(t)$ given by (2) stays between the two lines l_i and l'_i for all t in the knot point interval $[t_\kappa, t_\mu]$. By (18), it suffices to derive the condition for $Ap(t) \le d\ \forall t \in [t_\kappa, t_\mu]$.

First note that, by the definition of $B_k(t)$ in (4), $p(t)$ is written for unit knot point interval $[t_j, t_{j+1})$ as

$$p(t) = \sum_{i=0}^{k} \tau_{j-k+i} N_{i,k}(\alpha(t - t_j)),\ \ t \in [t_j, t_{j+1}), \tag{21}$$

and it depends on only the $k+1$ weights τ_{j-k}, τ_{j-k+1}, $\cdots$, τ_j. Introducing a new variable $u = \alpha(t - t_j)$, we see that $p(t)$ is written as $\hat{p}(u)$,

$$\hat{p}(u) = \sum_{i=0}^{k} \tau_{j-k+i} N_{i,k}(u),\ \ u \in [0, 1). \tag{22}$$

Thus if $A\tau_{j-k+i} \le d$ holds for $i = 0, 1, \cdots, k$, then using (5), we get

$$A\hat{p}(u) = \sum_{i=0}^{k} A\tau_{j-k+i} N_{i,k}(u) \le d \sum_{i=0}^{k} N_{i,k}(u) = d \tag{23}$$

or $Ap(t) \le d$ for all $t \in [t_j, t_{j+1})$. This interval can be readily extended to $[t_\kappa, t_\mu]$ by imposing the following condition,

$$A\tau_i \le d,\ i = \kappa - k, \kappa - k + 1, \cdots, \mu - 1. \tag{24}$$

Namely if (24) is satisfied, then it holds that

$$Ap(t) \le d\ \forall t \in [t_\kappa, t_\mu]. \tag{25}$$

Now the remaining task is to express the inequality condition (24) in terms of the control point vector $\hat{\tau}$ defined in (8). First rewrite (24) as

$$AT_{\kappa,\mu} \le \mathbf{1}^T_{\mu-\kappa+k} \otimes d \tag{26}$$

with $T_{\kappa,\mu} = [\ \tau_{\kappa-k}\ \tau_{\kappa-k+1}\ \cdots\ \tau_{\mu-1}]$ and $\mathbf{1}_i = [1\ 1\ \cdots\ 1]^T \in \mathbf{R}^i$. Then noting that $T_{\kappa,\mu}$ is a submatrix of τ consisting of its columns from $\kappa + 1$ through $\mu + k$, it can be expressed as

$$T_{\kappa,\mu} = \tau E_{\kappa,\mu}, \tag{27}$$

where $E_{\kappa,\mu} \in \mathbf{R}^{M \times (\mu-\kappa+k)}$ is defined by

$$E_{\kappa,\mu} = \begin{bmatrix} 0_{\mu-\kappa+k,\kappa} & I_{\mu-\kappa+k} & 0_{\mu-\kappa+k,M-\mu-k} \end{bmatrix}^T. \tag{28}$$

Thus (26) is written in τ as

$$A\tau E_{\kappa,\mu} \le \mathbf{1}^T_{\mu-\kappa+k} \otimes d. \tag{29}$$

Using a formula $\mathrm{vec}(AXB) = (B^T \otimes A)\mathrm{vec}X$ for matrices A, B, X of compatible dimensions (see e.g. (Lancaster and Tismenetsky, 1985)), and noting vec $\tau = \hat{\tau}$, (29) yields

$$(E^T_{\kappa,\mu} \otimes A)\hat{\tau} \le \mathbf{1}_{\mu-\kappa+k} \otimes d, \tag{30}$$

which is the desired expression.

For convenience we summarize the above developments as follows.

Proposition 1 *The trajectory $p(t)$ lies for all $t \in [s_i, s_{i+1}] = [t_\kappa, t_\mu]$ between the two lines l_i and l'_i if the control point vector $\hat{\tau}$ satisfies the following inequality.*

$$F_i \hat{\tau} \le h_i \tag{31}$$

where $F_i \in \mathbf{R}^{2(\mu-\kappa+k) \times 2M}$ and $h_i \in \mathbf{R}^{2(\mu-\kappa+k)}$ are given by $F_i = E^T_{\kappa,\mu} \otimes A$ and $h_i = \mathbf{1}_{\mu-\kappa+k} \otimes d$.

3.2 Centerline and Intermediate Time Instants

Recall that it is necessary to allocate the given time interval $[t_0, t_m]$ to n intervals $[s_i, s_{i+1}]$ with s_i being some knot point t_j for each $i = 0, 1, \cdots, n-1$. A natural way is to take the length of center line $C_i C_{i+1}$ (see Fig. 4) into account.

Let $C_i,\ i = 0, 1, \cdots, n$ be defined by

$$C_i = \gamma_i R_i + (1 - \gamma_i) L_i \tag{32}$$

for some γ_i ($0 \le \gamma_i \le 1$). Then for determining the time instants s_i in $[t_0, t_m]$, we employ the so-called centripetal distribution (Biagiotti and Melchiorri, 2008). Namely each s_i is determined so that the whole interval $[t_0, t_m]$ is divided into subintervals $[s_i, s_{i+1}]$ in proportion to the following value ζ_i,

$$\zeta_i = \|C_{i+1} - C_i\|^\nu \tag{33}$$

for $i = 0, 1, \cdots, n-1$, where $\nu (0 < \nu < 1)$ is usually taken as $\nu = 1/2$. Actually, each s_i is determined as a knot point t_j based on ζ_i. Also it is noted that the centripetal distribution method in above requires less accelerations than other method e.g. cord length distribution, distributed proportionally to $\|C_{i+1} - C_i\|$.

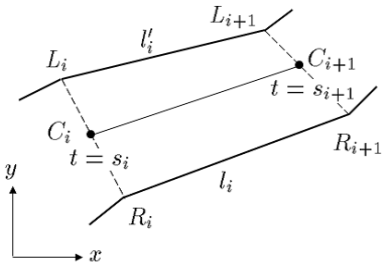


Figure 4. Centerline in the i-th quadrangle.

3.3 Smoothing Spline Trajectory

Obviously we require that, for all $i = 0, 1, \cdots, n-1$, the trajectory $p(t)$ lies between the two lines l_i and l_i' for all $t \in [s_i, s_{i+1}] = [t_\kappa, t_\mu]$. Thus by Proposition 1, we impose

$$F_i \hat{\tau} \le h_i,\ i = 0, 1, \cdots, n-1. \tag{34}$$

Usually an initial and terminal conditions of $p(t)$ are given, and typical examples are

$$p(t_0) = p_0,\ p^{(1)}(t_0) = 0,\ p^{(2)}(t_0) = 0 \tag{35}$$

$$p(t_m) = p_m,\ p^{(1)}(t_m) = 0,\ p^{(2)}(t_m) = 0 \tag{36}$$

with p_0 and p_m lying on the line segments R_0L_0 and R_nL_n respectively. Noting that $p(t)$ in (2) is written as $p(t) = \tau b(t)$ and hence $p(t) = \text{vec}\, p(t) = \text{vec}\,(\tau b(t)) = (b(t)^T \otimes I_2)\hat{\tau}$, the above initial and final conditions can be written in terms of $\hat{\tau}$ as follows.

$$H(t_0)\hat{\tau} = \bar{h}_0,\ H(t_m)\hat{\tau} = \bar{h}_m \tag{37}$$

where $H(t) \in \mathbf{R}^{6\times 2M}$, $\bar{h}_0 \in \mathbf{R}^6$ and $\bar{h}_m \in \mathbf{R}^6$ are

$$H(t) = \begin{bmatrix} b(t)^T \otimes I_2 \\ b^{(1)}(t)^T \otimes I_2 \\ b^{(2)}(t)^T \otimes I_2 \end{bmatrix}, \bar{h}_0 = \begin{bmatrix} p_0 \\ 0_2 \\ 0_2 \end{bmatrix}, \bar{h}_m = \begin{bmatrix} p_m \\ 0_2 \\ 0_2 \end{bmatrix}. \tag{38}$$

The matrices $H(t_0)$ and $H(t_m)$ can be easily set up, for example by using Table 1 when $k = 3$.

Finally in this section we consider the function $f(t)$ used for approximation of smoothing splines in (7). A natural choice will be a function constructed from the piecewise linear centerline $C_0C_1\cdots C_n$ introduced by (32). Specifically we employ a linear function of t in each section as shown in Fig. 4 as

$$f_i(t) = q_i t + r_i,\ t \in [s_i, s_{i+1}]. \tag{39}$$

Here $q_i, r_i \in \mathbf{R}^2$ are determined so as to satisfy $f_i(s_i) = C_i$ and $f_i(s_{i+1}) = C_{i+1}$. Thus the vector g in (11) is computed from

$$g = \sum_{i=0}^{n-1} \int_{s_i}^{s_{i+1}} b(t) \otimes f_i(t) dt \tag{40}$$

Now we are ready to formulate the trajectory planning problem by constrained smoothing splines with $J(\hat{\tau})$ in (9), the boundary inequality conditions in (34) and initial and terminal conditions in (37). Namely the problem is to minimize the cost function,

$$\min_{\hat{\tau}\in\mathbf{R}^{2M}} J(\hat{\tau}) = \frac{1}{2}\hat{\tau}^T G \hat{\tau} - g^T \hat{\tau} \tag{41}$$

subject to the constraints of the form

$$A_{eq}\hat{\tau} = d_{eq},\ A_{in}\hat{\tau} \le d_{in}, \tag{42}$$

where G and g are given in (10) and (11) respectively, A_{eq} and d_{eq} are formed as the collection of equalities (37), and A_{in} and d_{in} as collection of inequalities in (34).

Note that, in the case of (37), $A_{eq}\hat{\tau} = d_{eq}$ consists of 12 equality constraints in $2M$ unknowns $\hat{\tau}$. On the other hand, the total number of inequality constraints resulting from (34) is computed as $2M + 2(n-1)k$, with $2(n-1)k$ more than the number of unknowns.

This is a convex quadratic programing problem and can be solved numerically by using software tool as the function 'quadprog' in MATLAB. If necessary, other constraints may be introduced such as constraints on the magnitude of velocity or acceleration as long as all the constraints are consistent.

Finally, it is noted that, by our construction, the planned trajectory satisfies the following proposition.

Proposition 2 *The planned spline trajectory $p(t)$ is guaranteed to stay, for each $i = 0, 1, \cdots, n-1$, between the two lines l_i and l_i' as the right and left boundaries for all time $t \in [s_i, s_{i+1}]$. In particular, $p(t)$ is in the corner quadrangle formed from the four lines $l_i, l_i', l_{i-1}, l_{i-1}'$ at time $t = s_i$ for $i = 1, 2, \cdots, n-1$.*

4 Numerical Examples

Two examples of trajectory planning are considered, where we use cubic splines, namely we set $k = 3$ and $l = 2$ in (2) and (7). MATLAB function 'quadprog' is used for numerical solution of quadratic programming problems.

4.1 Path with Piecewise Linear Boundaries

We consider a path with the right and left corners R_i and L_i given as in Table 2.

Table 2. The coordinates R_i and L_i of right and left corners for $i = 0, 1, \cdots, n$ with $n = 9$.

	0	1	2	3	4	5	6	7	8	9
R_i	1	2	2	8	3	8	2	13	13	14
	0	2	9	9	6	4	0	0	8	9
L_i	0	1	1	12	6	11	6	12	12	14
	0	2	10	10	6	4	1	1	8	10

For this path, we construct a smoothing spline trajectory $p(t)$ in time interval $[t_0, t_m] = [0, 10]$. The initial and final conditions are as given in (35) and (36) with $p_0 = (R_0 +$

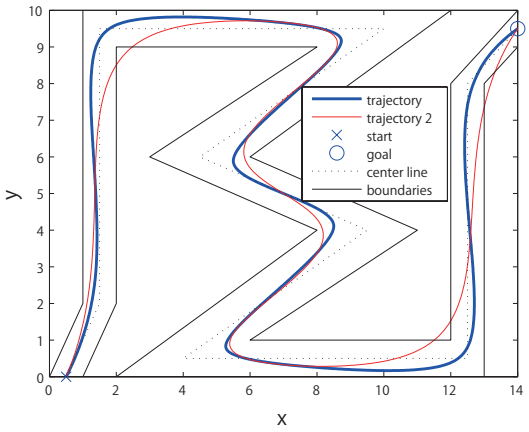


Figure 5. Constructed spline trajectories plotted in xy plane with and without the boundary constraints in (34).

$L_0)/2$ and $p_m = (R_9 + L_9)/2$. We set $\gamma_i = 1/2\ \forall i$ in (32) for centerline of the path, which together with (39) is used to set $f(t)$ in (7). The number of knot points $m = 80$, and hence the knot point interval is $t_{i+1} - t_i = 10/80 = 0.125$. The smoothing parameter is set as $\lambda = 0.01$ in (7).

Fig. 5 shows the given path with the boundaries denoted by black lines and the constructed trajectory $p(t) = [x(t),\ y(t)]^T$ on the xy plane in thick blue line. The start and goal positions are denoted by $\times$ and $\circ$ respectively, and the center line is shown in dotted lines. We observe that the trajectory satisfies the boundary constraints in addition to initial and final conditions. The red line is a smoothing spline trajectory $p2(t) = [x2(t),\ y2(t)]^T$ constructed similarly as $p(t)$ but without the inequality boundary constraints (34). It exceeds the boundaries at three corners, and the effect of introducing the inequality constraints is apparent.

4.2 Path in Obstacle Avoidance Problem

The proposed method can be used for planning trajectories of mobile robots in an environment with obstacles. As shown in Fig. 6, let us consider an environment with three 'obstacles', denoted by yellow rectangles, and the start and goal denoted by $\times$ and $\circ$ respectively. If we decide to take the upper route of the table-like obstacle in the figure, which will be shorter than taking the lower route, the path with piecewise linear boundaries can be defined, for example by setting the right and left corners (R_i, L_i) as shown in Table 3. These boundaries are shown in Fig. 6 in black dotted lines with the sign $\triangle$ and $\triangledown$ for right and left boundaries respectively.

We plan trajectories for this path with the time interval $[t_0, t_m] = [0, 10]$, initial and final conditions are as given in (35) and (36), $m = 50$ and the smoothing parameter $\lambda = 0.1$. Two cases of the parameters γ_i in (32) are considered for the centerline of the path as shown in Table 3. The first case (i) sets the line in the middle of the path, whereas the second case (ii) sets the line closer to the inner corner of the path. The planned trajectories are shown in the xy plane in thick blue line for Case (i) and thick red line for Case (ii). We see that reasonable trajectories have been generated where the second case seems more desirable, and that the parameters γ_i for the center line could be effectively used to adjust the trajectory. Note that the path is defined by only six pairs of right and left corners.

Table 3. The coordinates R_i and L_i of right and left corners and the parameter γ_i for centerline for $i = 0, 1, \cdots, n$ with $n = 5$.

	0	1	2	3	4	5
R_i	3 0	3 8	4 8	4 7	13 7	13 5
L_i	0 0	0 12	6 12	6 10	16 10	16 5
γ_i : Case (i)	1/2	1/2	1/2	1/2	1/2	1/2
γ_i : Case (ii)	1/2	2/3	2/3	1/3	2/3	1/2

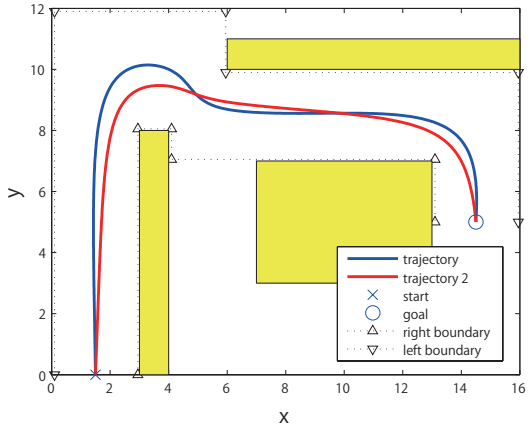


Figure 6. Constructed spline trajectories plotted in xy plane for two Cases (i) and (ii) of the centerline (see Table 3).

5 Concluding Remarks

We presented a method of trajectory planning for path with piecewise linear right and left boundaries. The trajectory is constructed as smoothing splines employing normalized uniform B-splines as the basis functions. Obviously, such boundaries can be described by simply providing a series of pairs of right and left corners (R_i, L_i), and the problem can be readily defined.

The boundary constraints could be expressed as linear inequality constraints on the control point vector $\hat{\tau}$. For constructing smoothing splines, we introduced a piecewise linear centerline of the path in accordance with each pair of right and left boundaries. An appropriate time interval for each piece of the centerline is given based on the centripetal distribution rule. It is shown that the problem is formulated as convex quadratic programming problem. We confirmed the effectiveness of the proposed method by two numerical examples.

One of the future issues is to relax the assumption (A1) in Section 3.1 to allow the multiple corner points as $R_i = R_{i+1}$. Extensions of present method are also important, e.g. to the cases of higher order boundary curves and to the planning in 3-dimensional space.

References

T. Berglund, A. Brodnik, H. Jonsson, M. Staffanson, and I. Soderkvist. Planning smooth and obstacle-avoiding b-spline paths for autonomous mining vehicles. *IEEE Trans. Automation Sci. and Eng.*, 7(1):167–172, 2010.

L. Biagiotti and C. Melchiorri. *Trajectory Planning for Automatic Machines and Robots*. Springer, 2008.

P. Crouch and J. Jackson. Dynamic interpolation and application to flight control. *J. of Guidance, Control and Dynamics*, 14: 814 – 822, 1991.

C. de Boor. *A practical guide to splines,Revised Edition*. Springer-Verlag, New York, 2001.

M. Egerstedt and C. F. Martin. Optimal trajectory planning and smoothing splines. *Automatica*, 37(7):1057–1064, 2001.

M. Egerstedt and C.F. Martin. *Control Theoretic Splines: optimal control, statistics and path planning*. Princeton University Press, New Jersey, 2010.

H. Fujioka and H. Kano. Optimal vector smoothing splines with coupled constraints. *Trans. Institute of Systems, Control and Information Engineers*, 25(11):299–307, 2012.

P. Gallina and A. Gasparetto. A technique to analytically formulate and to solve the 2-dimensional constrained trajectory planning problem for a mobile robot. *J. Intelligent and Robotic Systems*, 27:237–262, 2000.

H. Kano, H. Nakata, and C. F. Martin. Optimal curve fitting and smoothing using normalized uniform b-splines : A tool for studying complex systems. *Applied Mathematics and Computation*, 169(1):96–128, 2005.

H. Kano, H. Fujioka, and C. F. Martin. Optimal smoothing and interpolating splines with constraints. *Applied Mathematics and Computation*, 218(5):1831–1844, 2011.

H. Kano, H. Fujioka, and C. Martin. Optimal smoothing spline with constraints on its derivatives. *SICE Journal of Control, Measurement, and System Integration*, 7(2):104–111, 2014.

W. Khalil and E. Dombre. *Modeling, Identification and Control of Robots*. Hermes Penton Ltd., 2002.

J. Z. Kolter and A. Y. Ng. Task-space trajectories via cubic spline optimization. In *Proc. of the 2009 Int. Conf. on Robotics and Automation*, pages 1675–1682, Kobe, Japan, May 12-17, 2009.

P. Lancaster and M. Tismenetsky. *The Theory of Matrices, Second Edition*. Academic Press, 1985.

A. Piazzi and A. Visioli. Global minimum-jerk trajectory planning of robot manipulators. *IEEE Trans. Industrial Electronics*, 47(1):140 – 149, 2000.

S. Saska, M. Macas, L. Preucil, and L. Lhotska. Robot path planning using particle swarm optimization of Ferguson splines. In *Proc. IEEE Conf. on Emerging Technologies and Factory Automation (ETFA '06)*, pages 833 – 839, Prague, Sept. 20-22, 2006.

G. Wahba. *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics, 59, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

A Harvest Vehicle with Pneumatic Servo System for gathering a Harvest and its Simulation Study

Katsumi Moriwaki

Department of Mechanical Engineering, Daido University, Japan, moriwaki@daido-it.ac.jp

Abstract

A series of works such as harvesting and transporting in a farm is one of such works with so care as not to damage the harvest in order to maintain the value of harvests. We are developing an autonomous cart for gathering a harvest with the bed to be controlled to keep in horizontal level at work and in transit, in order to avoid harvests gathered in particular area of the bed and to keep away from being damaged in harvests. It is proposed a method of autonomous steering control of a harvest vehicle with maintaining the horizontal level of the bed of the cart with air cylinder suspension systems. It is shown that the problem of a level control of vehicle's bed can be formulated as one of optimal control problems. Finally, its simulation study is considered.

Keywords: level control of vehicle's bed, pneumatic servo control, autonomous vehicle, terrain farm, harvesting

1 Introduction

The work of harvesting of the fields is one of tough works which need the efficiency of materials handling, while it is necessary for guarantee of commodity value to treat crops carefully so that a crack may not be attached to crops. A harvest cart is, therefore, one of necessary apparatuses for both a professional farmer and a urban farmer to lighten their work load to harvest the fields. There are many harvest carts developed and sold in various size, from smalled-sized to mega-sized. In this research, it aims at development of the conveyance cart which can carry a harvest without swaying of the loading bed by performing level surface maintaining control, in order for the crops not to be damaged. We proposed the basic structure for level control of a harvest bed using pneumatic system (Moriwaki, 2012). It was proposed the harvest vehicle with a level-controlled harvest bed and considered the optimal control problem of a level control of a harvest bed (Moriwaki, 2013, 2016). We, furthermore, consider to realize the system of intelligent harvesting which performs autonomous collection of crops in a cultivated land and autonomous carrying them to the crops shed without deterioration of the commodity value of them. Figure 1 shows an examples of harvest vehicle, which carries crops from the cultivated land to the harvest gathering place. The harvest cart often jolts over the rough cultivated land, and crops vibrate on the loading bed of the harvest cart. Crops collide together or with the wall of a harvest stand, It damages the surface of crops and reduces the commodity value of crops. If crops without cracks can be harvested and shipped, they can be sold for a high price at a market. We have started a development program of a harvest vehicle which can perform harvesting operation and maintain the level surface of its loading bed, and can autonomously go to a harvest collection place.

The suspension model of a vehicle is considered in Section 2. After considering the structure of the loading bed of a harvest vehicle in Section 3, A model of level controlled bed by a pneumatic servo system is proposed in Section 4. The control system of a harvest bed with a pair of pneumatic cylinder is proposed in Section 5, where is also shown some numerical simulations and experimental study. Finally, it is considered the problem of level control of harvest's bed on the terrain farm land in Section 6.

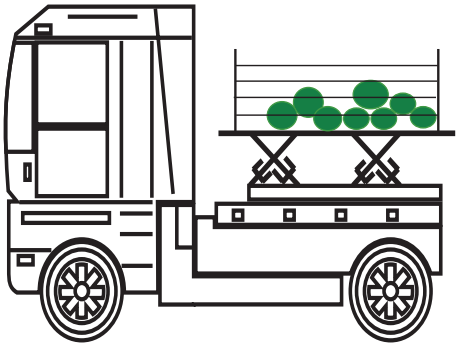


Figure 1. A harvest vehicle with active pneumatic suspension bed.

2 Suspension model of a vehicle dynamics

The control problem of leveling the load bed of a harvest vehicle is deeply relation to the vertical dynamics of a vehicle, its suspension dynamics. The features of car vertical dynamics in (X-Z)-plane are described by Figure 2 (Abe, 1992; Andrezejewski and Awrejcewicz, 2005) with active suspension system in which the external control forces is used to suppress the uncomfortable bouncing motion and pitching motion.

The equations of motions for two degrees of freedom in (X-Z)-plane with the constant velocity V are

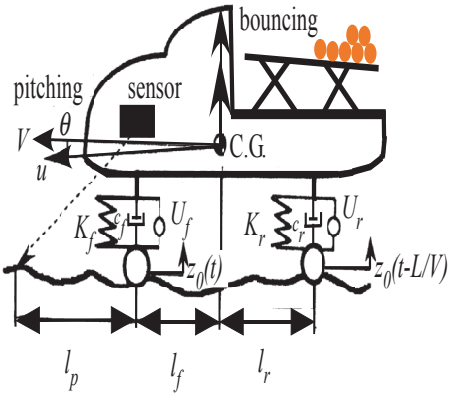


Figure 2. Active suspension control of a harvest vehicle.

$$\text{bouncing motion} : m\ddot{z} = F_f + F_r \tag{1}$$

$$\text{pitching motion} : I_y\ddot{\theta} = -l_f F_f + l_r F_r \tag{2}$$

where m is the mass of a vehicle, z is the vertical displacement of the center of gravity (CG) of a car(Figure 2), I_y is the moment of inertia around y-axis, θ is the pitching angle of the center of gravity (CG) of a car and the external forces acting on a front wheel and the rear wheel from the road surface are written by F_f and F_r, respectively. The state space model of the vertical dynamics is derived from (1) and (2), where it is assumed that the vehicle is affected from the terrain of the farm land as the unknown input U_f, U_r and the surface roughness $z_{0f} = z_0(t)$, $z_{0r} = z_0(t - L/V)$, where $L = l_f + l_r$.

$$\begin{bmatrix} \dot{z} \\ \dot{\theta} \\ \dot{w} \\ \dot{q} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \tilde{a}_{31} & \tilde{a}_{32} & \tilde{a}_{33} & \tilde{a}_{34} \\ \tilde{a}_{41} & \tilde{a}_{42} & \tilde{a}_{43} & \tilde{a}_{44} \end{bmatrix} \begin{bmatrix} z \\ \theta \\ w \\ q \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \tilde{b}_{31} & \tilde{b}_{32} \\ \tilde{b}_{41} & \tilde{b}_{42} \end{bmatrix} \begin{bmatrix} U_f \\ U_r \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \tilde{h}_{31} & \tilde{h}_{32} \\ \tilde{h}_{41} & \tilde{h}_{42} \end{bmatrix} \begin{bmatrix} z_{0f} \\ z_{0r} \end{bmatrix} \tag{3}$$

where $w := \dot{z}$, $q := \dot{\theta}$ and

$$\left.\begin{aligned} \tilde{a}_{31} &= -\tfrac{K_f + K_r}{m}, \quad \tilde{a}_{32} = \tfrac{K_f l_f - K_r l_r}{m} \\ \tilde{a}_{33} &= -\tfrac{C_f + C_r}{m}, \quad \tilde{a}_{34} = \tfrac{C_f l_f - C_r l_r}{m} \\ \tilde{a}_{41} &= \tfrac{K_f l_f - K_r l_r}{I_y}, \quad \tilde{a}_{42} = -\tfrac{K_f l_f^2 + K_r l_r^2}{I_y} \\ \tilde{a}_{43} &= \tfrac{C_f l_f - C_r l_r}{I_y}, \quad \tilde{a}_{42} = -\tfrac{C_f l_f^2 + C_r l_r^2}{I_y} \end{aligned}\right\} \tag{4}$$

$$\left.\begin{aligned} \tilde{b}_{31} &= \tfrac{1}{m}, \quad \tilde{b}_{32} = \tfrac{1}{m} \\ \tilde{b}_{41} &= -\tfrac{l_f}{I_y}, \quad \tilde{b}_{42} = \tfrac{l_r}{I_y} \end{aligned}\right\} \tag{5}$$

$$\left.\begin{aligned} \tilde{h}_{31} &= \tfrac{K_f}{m}, \quad \tilde{h}_{32} = \tfrac{K_r}{m} \\ \tilde{h}_{41} &= -\tfrac{K_f l_f}{I_y}, \quad \tilde{b}_{42} = \tfrac{K_r l_r}{I_y} \end{aligned}\right\} \tag{6}$$

where

K_f : stiffness coefficient for front suspension
K_r : stiffness coefficient for rear suspension
C_f : damping coefficient for front suspension
C_r : damping coefficient for rear suspension

3 Structure of the loading bed of a harvest vehicle

It is considered that the structure of the loading platform with the level control mechanism using pneumatic cylinders of an autonomous harvest cart in this section. We propose a control system which maintains a loading platform horizontally while harvested crops is gathered on the loading platform. When crops are taken in and it is stored by the loading platform, there are often for crops to be thrown into the harvest cart and stored randomly on a loading-platform. If the level maintaining of a loading-platform becomes difficult according to the random loading of crops, a loading-platform inclines and crops may roll in the direction of a lower side, then the damage to the crops caused by collision with other crops or with the wall of the loading-platform may occur, and it may produce deterioration of the commodity value of crops. Moreover, if crops are loaded to the specific side of a loading platform and a loading-platform inclines, the efficiency carrying crops will be affected. In order to prevent such a bad influence to an agricultural harvest work, this research considers a method of the level control of the loading bed of an autonomous harvest cart.

There are many cases in position the load bed with crops (Figure 3). Cases (a) and (b) are situations of the bed on flat field, the bed in (a) is maintained horizontally with the harvest orderly, on the other hand, the bed (b) leans with gathered crops on one side. Case (c) and (d) are situations of the bed on terrain field, the bed in (c) leans because of gathered harvest and uneven surface of the field, our aim is to propose the controlled bed to maintain horizontally, whether on uneven field or with gathered crops in ether side. In this paper, it aims at realizing the level control of a loading-platform by using position control of pneumatic cylinders with electromagnetic pressure proportional valves and a computer (Figure 4). In Figure 4, the load f_L, which is induced for crops to drop on the load platform, presses down the cylinder head, whose mass is denoted by M, and the loading platform leans down to either side. The deviation is detected by a height sensor and

is put into a computer as the reference. The computer calculates the control input e, with which a pressure valve controls the cylinder pressure so that the loading platform move back to the horizontally.

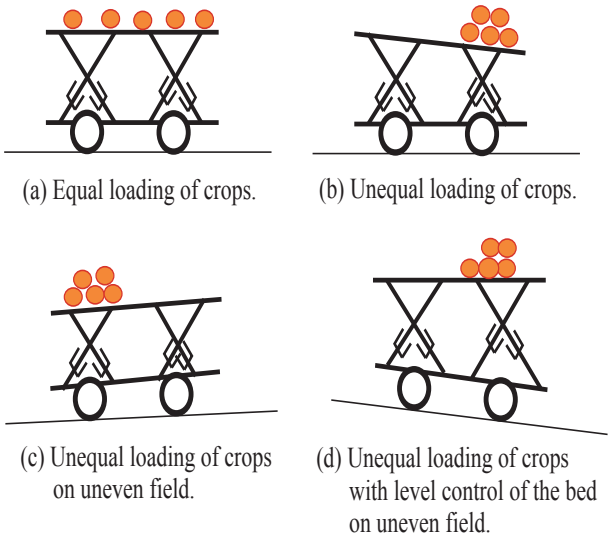


Figure 3. Harvest quality managing cart.

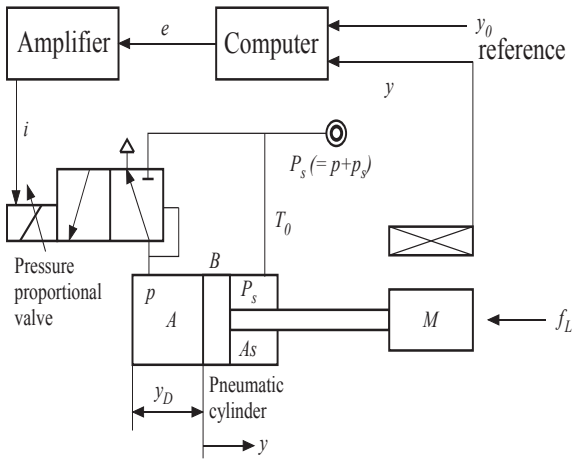


Figure 4. Position control of a pneumatic cylinder with a pressure proportional valve.

4 Modeling of a pneumatic cylinder with a pressure proportional valve

A pneumatic cylinder servo control system is expected to be applied to various automation systems, because it has some advantages such as high power/weight ratio, functions of impact absorption and rigorous force control owing to air compressibility. However, the compressibility makes it a high order and nonlinear system, its exact modeling and parameter estimation are not easy (Noritsugu and Takaiwa, 1993; Song, et al., 1997). There are many paper published about pnematic servo system up to now, almost all of them are, however, dealed with the problem of the position control or force control of a simple pneumatic cylinder. There is few study about the practical control system with the pneumatic servo system (Moriwaki, 2012, 2013, 2016).

4.1 A model for a pneumatic cylinder

The air mass flow rate W [kg/s] which flows into a cylinder can be expressed as follows.

$$W = \frac{1}{RT_0}\{p\dot{V} + (\frac{V}{\kappa})\dot{p}\} \tag{7}$$

where

$$V = A(y_D + y) \tag{8}$$

and A is a area of the cylinder $[m^2]$, T_0 is an absolute temperature $[K]$, κ is the ratio of specific heat of the air (≈ 1.4) and R is the gas constant $(= 287m^2/s^2K)$.

The motion of a rod of the pneumatic cylinder is described by the following equation:

$$M\ddot{y} + B\dot{y} = Ap - A_s p_s + f_L \tag{9}$$

where $y, A_s, p, p_s, M, B, f_L$ are the position of rod in a cylinder, a area of rod-side, a nominal portion of the supply pressure P_s, a fluctuation portion of the supply pressure P_s, a inertial mass of the cylinder, a viscous coefficient of friction of the cylinder and a external force, respectively. It is usually assumed that a fluctuation portion of the supply pressure is sufficiently small, then we can put $p_s = 0$ in (9). Therefore, we obtain the following equation from (9).

$$M\ddot{y} + B\dot{y} = Ap + f_L \tag{10}$$

4.2 A model for a pressure proportional valve

It is well known that the transient characteristics of a pneumatic cylinder is remarkably affected to a control flow, inertial mass, death volume of a valve (Tanaka, 1987; Kagawa and Cai, 2010). The servomechanism which controls the cylinder of a single rod type by the three direction type of a pressure proportional valve is used in the proposed system.

The mass flow rate W_v may be linearized with respect to the displacement x_v of a valve, then we obtain the following equation (Tanaka, 1987; Kagawa and Cai, 2010):

$$W_v = (\frac{\partial W_v}{\partial x_v})x_v + (\frac{\partial W_v}{\partial p})p = k_v x_v - k_p p \tag{11}$$

where $k_v, k_p,$ are a flow gain, a pressure flow coefficient $(k_p > 0)$, respectively. Accompanying the motion of a cylinder to the mass flow rate W_v given by (11), the energy equilibrium equation is given by

$$W_v + \frac{1}{RT_0}A_v p\dot{x_v} = W \tag{12}$$

Equation (12) is transformed and the following (13) is obtained.

$$W_v = \frac{1}{RT_0}(pA\dot{y} + \frac{V}{\kappa}\dot{p} - A_v p\dot{x_v}) \tag{13}$$

where A_v is the area of a feedback pressure of a valve.

For an electromagnetic force F_c of a solenoid, the equation of motion of a pressure valve can be expressed as follows (Tanaka, 1987).

$$m\ddot{x_v} + b\dot{x_v} = F_c - A_v p \tag{14}$$

where m is the inertial mass of a moving rod of a cylinder and b is the coefficient of viscous damping.

5 The position control system of a pneumatic cylinder using a pressure proportional valve

5.1 A block diagram of a pneumatic servo system

The solenoid's electromagnetic force F_c can be approximated by the following equation, denoting feedback gains of a cylinder position y and its moving velocity $\dot{y}$ by g_{fs}, g_{fv}, respectively,

$$F_c \cong \frac{a_1(a_0 y_0 - g_{fs}y - g_{fv}\dot{y})}{\tau_0 s + 1} \tag{15}$$

where a_1, a_0, τ_0 denote the electromagnetic force conversion coefficient, the voltage conversion coefficient w.r.t. reference input y_0, and the time constant of the solenoid.

Combining (10), (11) and (13)–(15), then the position control system is obtained by Figure 5, in which a cylinder position y can be maintained at the reference position y_0 under an unexpected load (or force) being added to the cylinder. In Figure 5, time constants $\tau_i\,(i = 1, \cdots, 4)$ denote $\tau_1 = \frac{k_c}{k_p}$, $\tau_2 = \frac{m}{b}$, $\tau_3 = \frac{A_v}{k_x}$, $\tau_4 = \frac{M}{B}$, respectively, and k_x is the flow gain of the valve, k_p is the pressure flow coefficient, and k_c is the compliance of air pressure. In the usual air pressure servomechanism, $\tau_0, \tau_2, \text{and} \tau_3$ may be very small and can be ignored (Tanaka, 1987). The time constant τ_4 may be changed to $(\tau_4 + \Delta)$ depending on the loaded mass.

5.2 A reduced block diagram of a pneumatic servo system

If the pneumatic servo system has the condition $\tau_4 > \tau_1$, which means its inertia characteristics are dominant than its volume characteristics in the system, the block diagram shown in Figure 5 can be reduced as Figure 6. In Figure 6, the parameters $\omega_M, \zeta_M, k_{Lp}$ denote the following values respectively.

$$\omega_M = \sqrt{\frac{\{A_v + (a_1 g_{fv} A)/B\}k_x}{bk_p\tau_4}} \tag{16}$$

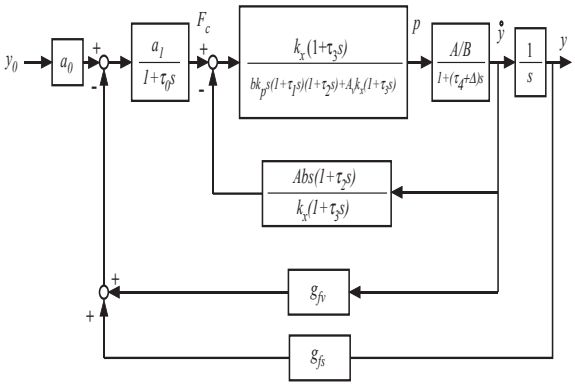


Figure 5. The position control system by a pneumatic cylinder.

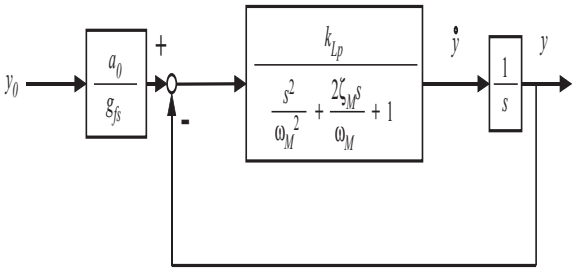


Figure 6. The reduced position control system by a pneumatic cylinder.

$$\zeta_M = \frac{b(k_p + A^2/B) + A_v k_x \tau_4}{2\sqrt{bk_v\tau_4 k_x\{A_v + (a_1 g_{fv} A)/B\}}} \tag{17}$$

$$k_{Lp} = \frac{(a_1 g_{fp} A)/B}{A_v + (a_1 g_{fv} A)/B} \tag{18}$$

The characteristics of step responce of the pneumatic servo system is analyzed theoretically and the parameter k_{Lp}/ω_M is necessary to be set 0.4 or less in order to obtain the undershoot step response (Tanaka, 1987).

5.3 A state space model for leveling control of the loading bed of a harvest cart

The loading bed of a harvest cart, which we proposed in this paper, is controlled by a pair of pneumatic cylinders as shown in Figure 7, which is appeared in later section. The structure of dynamics of the proposed level control system is shown in Figure 8.

The rods of pneumatic cylinder (a) and (b) are connected to the loading bed with rotation and bottoms of cylinders are connected to the base of the harvest cart with the angle θ of inclination. The angle θ is assumed to be fixed, i.e. $\sin\theta$ is to be constant. The height z_a, z_b of the both ends of the loading bed are controlled by the position of the rod y_a, y_b, respectively, as

$$z_a = y_a \sin\theta, \;\; z_b = y_b \sin\theta \tag{19}$$

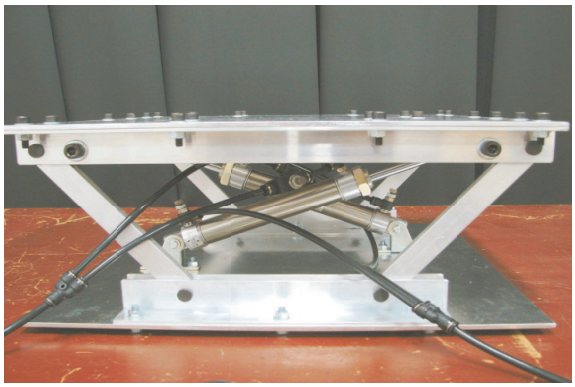


Figure 7. The experimental device for level control of the loading platform.

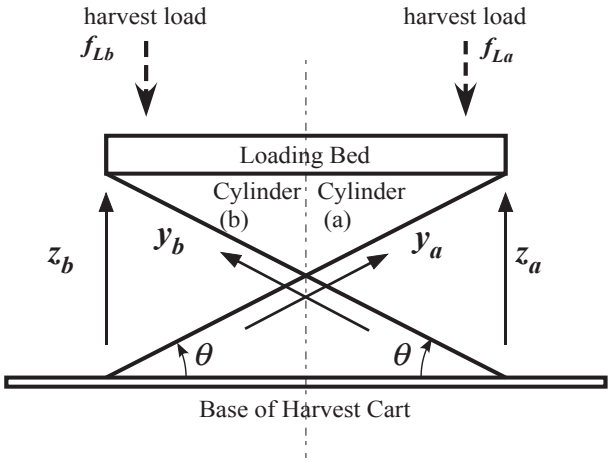


Figure 8. The structure of a dynamic model for level control of the loading bed.

with the state space model as follows: $(i = a, b)$

$$\frac{d}{dt}\begin{bmatrix} y_i \\ \dot{y}_i \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -\frac{2B}{M} \end{bmatrix}\begin{bmatrix} y_i \\ \dot{y}_i \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{2A}{M} \end{bmatrix} p_i + \begin{bmatrix} 0 \\ \frac{2A}{M} \end{bmatrix} f_{Li}, \tag{20}$$

where p_i is the control input and f_{Li} is the harvest load. The objective of level control of the loading bed is, therefore, maintain the difference $z_a - z_b$ to be minimized. That is the difference v defined by

$$v = \sqrt{(z_a - z_b)^2} = \sqrt{(y_a - y_b)^2}\sin\theta \tag{21}$$

to be reduced as soon as possible.

5.4 Simulation results of the level control system

The harvest is loaded on the loading bed (Figure 8) of the harvest cart unevenly from the farm. We have simulated the level control of the loading bed under (a) the periodical loading from the both ends with different weights and a phase difference and (b) the random impulsive loading from the both ends with different weights and a phase difference. Figure 9 shows the external harvest loadings f_{La}, f_{Lb} and the difference v of the loading bed (21) for the case (a). Figure 10 shows the external harvest loadings f_{La}, f_{Lb} and the difference v of the loading bed (21) for the case (b). In Figures 9, 10 amplitudes of f_{La}, f_{Lb}, v are normalized and scaled.

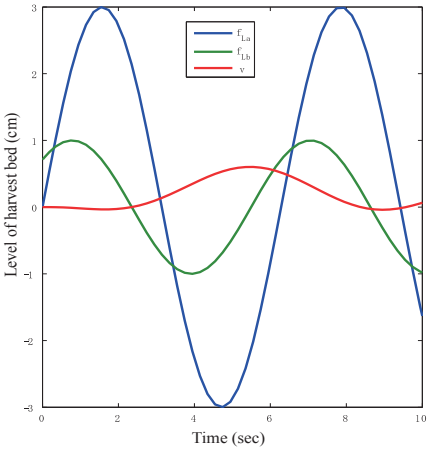


Figure 9. A state of level of the harvest bed under periodical loading.

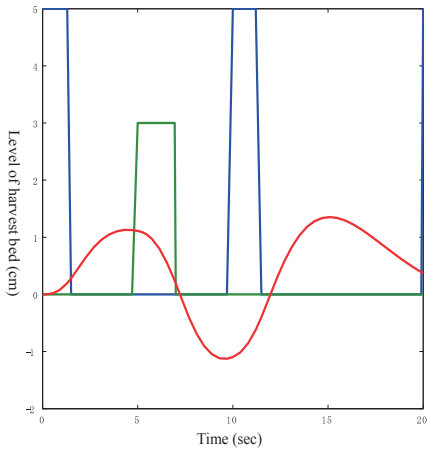


Figure 10. A state of level of the harvest bed under random impulsive loading

6 Leveling control of the harvest bed on the terrain farm land

The harvest vehicle usually moves on rough terrain to carry crops to the cargo-pickup point at the farm. In this section, we consider the control problem to maintain the level of a harvest bed under the existence of pitching motion of the suspension. Let's assume there is a pitching

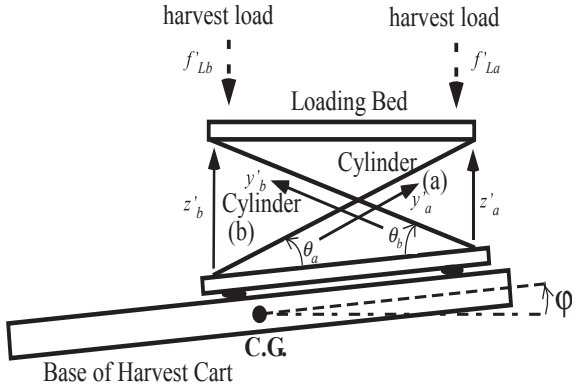


Figure 11. The model of a harvest bed on the platform of the vehicle on rough terrain.

angle φ, while the harvest vehicle moves. Then the model of a harvest bed under pitching motion is shown in Figure 11.

The height z'_a, z'_b of the both ends of the loading bed are controlled by the position of the rod y_a, y_b, respectively, as

$$z'_a = y'_a \sin\theta_a, \;\; z'_b = y'_b \sin\theta_b \tag{22}$$

with the state space model as follows: $(i = a, b)$

$$\frac{d}{dt}\begin{bmatrix} y'_i \\ \dot{y}'_i \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -\frac{2B}{M} \end{bmatrix}\begin{bmatrix} y'_i \\ \dot{y}'_i \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{2A}{M} \end{bmatrix} p_i + \begin{bmatrix} 0 \\ \frac{2A}{M} \end{bmatrix} f'_{Li}, \tag{23}$$

where p_i is the control input and f'_{Li} is the harvest load. There is the incline of the base of a harvest vehicle, caused by the pitching angle φ. Therefore, we have to compensate for the incline in order to control the level of the harvest bed. The difference between z'_a and z'_b is

$$\Lambda = l_r \sin\varphi \tag{24}$$

where l_r is the distance between the center of gravity (C.G.) and the rear axle of the vehicle (Figure 2).

The objective of level control of the loading bed for this case is, therefore, maintain the difference $z'_b - z'_a$ to be equalized to the incline height Λ, as soon as possible. That is the difference $|\Lambda| - |v'|$ defined by

$$v' = \sqrt{(z'_a - z'_b)^2} = \sqrt{(y'_a \sin\theta_a - y'_b \sin\theta_b)^2} \tag{25}$$

to be reduced as soon as possible. It is not clear in this general case for the solution to level control of the harvest bed of the cart.

7 Conclusions

The autonomous harvest cart with level control of loading platform has been proposed in this paper. Crops can be stored in an uneven cultivated land so that reduction in commodity value may not be produced by using the proposed harvest cart. The pneumatic cylinder servo control system is used to control the level of loading platform. The proposed harvest cart is designed so that it can maintain the level of loading platform while loading of crops and following a bumpy road from a farmland.

References

M. Abe. *Vehicle Dynamics and Control*. Saikai-do Press, Tokyo, 1992.

R. Andrezejewski and J. Awrejcewicz. *Nonlinear Dynamics of a Wheeled Vehicle*. Springer, New York, 2005.

Toshiharu Kagawa and Maolin Cai. *Measurement and control of compressive fluid -An introduction to air pressure analysis-*. Japanese Industrial Publishing Company, Tokyo, 2010.

Katsumi Moriwaki. An autonomous cart for gathering a harvest and its control. *In Proc. of 2012 ISCIE Conference*, pages 35–36. ISCIE, Osaka, 2012.

Katsumi Moriwaki. Level control for the Loading Bed of a Harvest Cart by the Pneumatic Servo System and its Simulation Study. *In Proc. of 2013 8th EUROSIM*, pages 90–94. SIM, Cardiff, UK, 2013. doi:10.1109/EUROSIM.2013.26.

Katsumi Moriwaki. A harvest vehicle with pneumatic servo system for gathering a harvest and its control. *ScienceDirect*, 49-21:460–466. ELSEVIER, 2016.

M. Nagai, A. Okada, K. Komoridani, Y. Suda, K. Tani, H. Amijima, S. Nakajiro, H. Harada, M. Miyamoto, and H. Yoshioka. *Dynamics and Control of Vehicles*. Yoken-do Press, Tokyo, 1999.

Toshiro Noritsugu and Masahiro Takaiwa. Robust control of a pneumatic servo system using disturbance observer. *Trans. of SICE*, 29:86–93. SICE, Tokyo, 1993.

Toshiro Noritsugu and Masahiro Takaiwa. Positioning control of pneumatic servo system with pressure control loop using disturbance observer. *Trans. of SICE*, 31 :1970–1977. SICE, Tokyo, 1995.

Junbo Song, Kazunori Kadowaki and Yoshihisa Ishida. Practical model reference robust control for pneumatic servo system. *Trans. of SICE*, 33:995–1001. SICE, Tokyo, 1997.

Hirohisa Tanaka. *Digital control of a hydraulic system and a pneumatic system and its applications*. Kindai Tosho Press, Tokyo, 1987.

Creating Social-aware Evacuation Plans based on a GIS-enable Agent-based Simulation

Kasemsak Padungpien and Worawan Marurngsith
Department of Computer Science, Thammasat University, Thailand, wdc@cs.tu.ac.th

Abstract

In disaster preparedness, agent-based simulation (ABS) is an effective tool for aggregating information on evacuees affected by disasters. An agent usually imitates a household; and its actions are normally specified by decision models based on risk perception and social elements. Assigning socially connected households to the same shelters can better utilise resources. However, by pre-assigning specified regions to shelters, social connections are often omitted when developing evacuation plans at policy level. Thus, this paper presents a method to create social-aware evacuation plans. A GIS-enabled ABS is used to estimate evacuation demand and group evacuees according to their social connection data prior to assign them to the nearest shelters. The evacuation plans generated by the proposed method are compared against the travel-cost optimisation plans solved by using a linear model. The results obtained from a case study show that the social-aware evacuation plans offer a slightly better utilisation of shelter capacity; take similar time to evacuate households; yet could save nearly three hours to achieve complete evacuation. These results seem to confirm the competitiveness of social-aware evacuation plans as an option for evacuation planning at policy level.

Keywords: agent-based simulation, social connection, K-means clustering, linear programming, evacuation and shelter planning, GeoMASON

1 Introduction

Evacuation and shelter planning is essential to decrease the severity of loss and damage caused by either natural or man-made disasters. Evacuation studies may be classified into three groups: disaster models, traffic models, and traffic disaster evacuation models (Bae et al., 2014). Previous studies from 1985 – 2014 as reviewed in (Bae et al., 2014; NaBanerjee, 2015) confirmed that agent-based simulation (ABS) is the most popular tool used for gaining aggregated information of an evacuation. A finding has pinpointed that an evacuation decision is determined by a combination of household characteristics and by both capacity-related and hazard-related factors (Lim et al., 2016).

The first step in evacuation planning, is to forecast the number of people to be evacuated and the time to start evacuation, *i.e*. the demand estimation. ABS models were successfully used in (FlötterödLämmel, 2010; NaBanerjee, 2015; Yin et al., 2014) to gain insight into the evolution of evacuation demand when considering both risk factors and social elements. Proactive agents, *i.e.* goal-oriented agents searching for safe places or performing tasks to prepare for evacuation, were used to model households (NaBanerjee, 2015), individuals (FlötterödLämmel, 2010) and patients (Yin et al., 2014). The agents take actions based on the decision models integrated into a simulation (Gama et al., 2016; Lei et al., 2012; Lim et al., 2016; Yin et al., 2014). These models are mainly probabilistic decision models derived from human behaviour in response to hazards and evacuees' demographic characteristics (e.g. sex, age, education, social connection, and economical status) collected from historical records and interviews.

Evacuation plans obtained from aggregated results of ABS are not only determined by the shortest routes, but also by the behaviour of households based on their social connections (*i.e.*, how friends and relatives might influence a household). Research efforts have successfully integrated some social connection factors into decision models as an attempt to better understand dynamic factors in an evacuation decision (Lim et al., 2016; Yin et al., 2014). Their findings confirmed the role of social connections among evacuees in effective emergency response, as described in (Hofinger et al., 2014; Wachtendorf et al., 2013). Aggregating agents' social interaction using ABS may not be applicable when creating plans based on regional segmentation. Besides, shelter assignment techniques based on regions are usually focussed on transportation optimisation rather than social connections. Consequently, social connections are often omitted when developing evacuation plans at policy level.

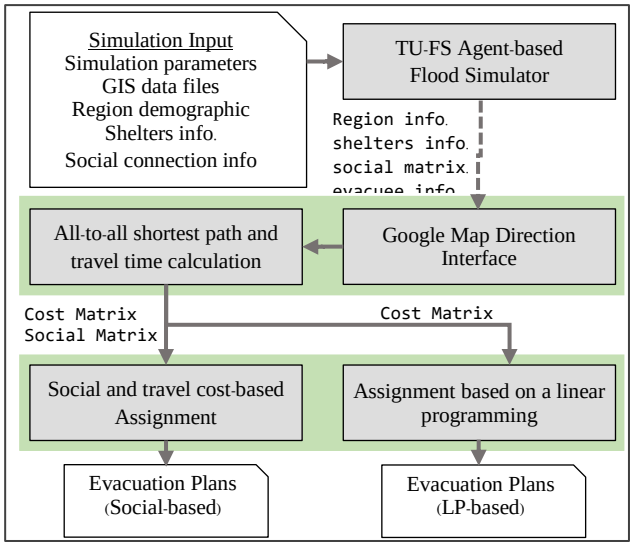


Figure 1. Overall process to generate the evacuation and shelter plans.

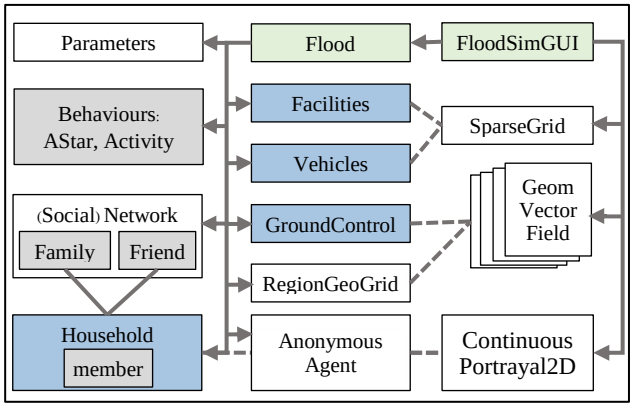


Figure 2. Classes and components in the TU-FS simulator.

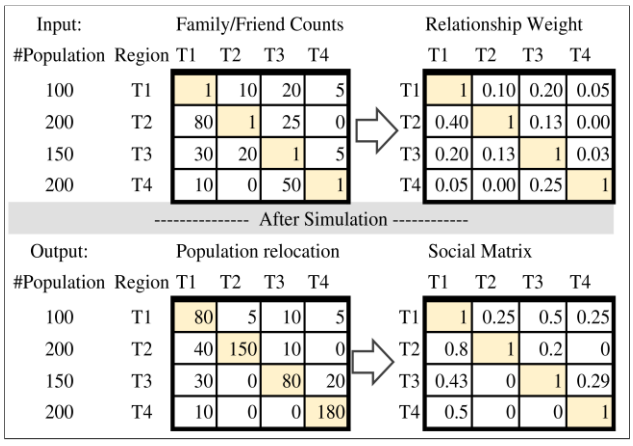


Figure 3. Example of the social connection and the social matrix.

This paper presents a method to create a social-aware evacuation plan. A GIS-enabled ABS called the TU-FS simulator (Vijitpornkul and Marurngsith, 2015), implemented on the GeoMASON framework (George Mason University, 2013), is used to estimate evacuation demand and obtain social connection data. These data are used to classify evacuees into groups and assign them to the nearest shelters. This paper makes three contributions.

(1) An estimation of evacuation demands and households' social connections using a GIS-enabled agent-based flood simulator is demonstrated.

(2) A shelter assignment method to generate social-aware evacuation plans by using the K-means clustering and realistic traffic information obtained via the Google Map Direction API (Google Developer, 2016) is proposed.

(3) A linear optimisation model for solving the capacity-aware shelter assignment problem based on minimising travel-cost is presented. Evacuation plans generated by this model were compared against the social-aware plans and this confirmed the effectiveness of the proposed social-aware method.

A case study, covering the central part of Lop Buri province in Thailand, was conducted to demonstrate the applicability of the proposed method. The results confirm the competitiveness of social-aware evacuation plans as an appropriate alternative to current policies for evacuation.

The rest of this paper is organised as follows. The next section introduces an agent-based flood simulation used in this study. Section 3 presents the proposed evacuation and shelter planning method. Section 4 analyses the efficiency gained from allocating ten shelters with the proposed method in comparison to the linear optimisation model. Finally, Section 5 brings this paper to a conclusion.

2 GIS-Enabled Agent-Based Flood Simulation

The overall process to generate evacuation and shelter plans is shown in Figure 1. Firstly, users must setup the flood scenarios by providing six inputs *i.e.*, (1) the simulation parameters, (2) the GIS data files, (3) information about the affected regions, (4) information about shelters and other facilities, (5) social connection information and (6) the household demographics. All these inputs are then passed to the TU-FS simulator to run a simulation. The TU-FS simulator gathers the amount of population in the affected regions and processes it into four output files that represent the evacuation demand. These four files contain information about the affected regions, shelters, households (evacuees), and a social matrix, respectively. Data contained in those four files is processed using the Google Map Direction interface followed by an all-to-all shortest path and travel time calculation. The process creates two matrices, namely: a cost matrix, and a social matrix. In our approach, we use both matrices to assign shelters and generate evacuation plans. In contrast, the linear model only uses the cost matrix to deal with shelter-assignment and to produce evacuation plans.

We have modified the original TU-FS simulator to be able to simulate pre-evacuation activity in different scenarios. The core simulation model and the graphical user interface of the model were implemented in two Java classes: Flood and FloodSimGUI (Figure 2). The Flood class contains objects from four `steppable` classes (`Facility`, `Vehicles`, `GroundControl`, and `Household`), four custom classes (`AStar`, `Activity`, `Family`, `Friend`) and other objects inherited from the MASON and GeoMASON libraries. The `steppable` classes implement objects whose behaviour is updated at every time step during a simulation. Custom classes are used to implement specific behaviours of the objects. For example, the `AStar` class implements an algorithm to assign a household or a vehicle to a certain travel path. Four objects extended from the Mason library are used to enable customised parameters. These are the `ContinuousPortrayal2D`, `SparseGrid`, `Network`, and `Param`. These objects are also used for creating both the social network relationship among households and the layered environments in the model. GIS data files are added into the model using two classes of the GeoMASON library, *i.e.*, the `GeomVectorField` and `RegionGeoGrid` objects. Details of the MASON simulation engine can be found in (Luke, 2015).

The `GroundControl` class manages the GIS data layers in the model. To do this, three parameters are required, namely: a list of GIS ASCII format files, the affected region and shelter location. The GIS layers of an area under study are read from the list of GIS ASCII files and placed on top of the 2D continuous space containing the households' agents. These layers represent the political boundary of a region, the ground water, the transportation (road network), and the flood map. The flood map layer consists of a set of files that are named in sequence. That map is updated at every specified simulation time step (*e.g.* every simulated hour). The TU-FS simulator uses the WGS84 (EPSG4326) coordinate reference system. The list of regions and shelter locations is stored in a CSV and fed into the simulation by the `GroundControl` class. The GPS locations are also sent to the `Facilities` class to create agents for shelters and sub-districts.

Parameters related to household demographics and social connections are also specified in the CSV-format files. The household demographics file records the number of affected regions or sub-districts under study. That information is kept in rows, where every row specifies information about the households in a sub-district as the percentage of the population having certain attributes. For example, these attributes could be personal characteristics (sex, height, average age), health status (healthy, infected, sick, dead), number of families, mobility state. The parameters used to derive social connections are in the `FriendInfo` and `FamilyInfo` files. These files contain 2D matrices that represent a Count and a Size part. Each part of the file has a 2D matrix of R×R size, where R is the number of the affected regions. Figure 3. (top left) depicts the structure of the Count part of the file. The Size part is kept in a matrix where each cell (in row i column j) represents the average of friend/family members that live in sub-district j of a person that lives in sub-district i. The `FriendInfo` and `FamilyInfo` files are used to create Household, Family, Friend and Vehicle objects to meet the specified parameters. Social network relationships are set using the MASON's social network facility.

Once all objects have been successfully initialised, the TU-FS simulator runs a simulation until either a defined step or a trigger marked in the flood map is reached. At every step in the simulation, household agents may move to join with friends or family. The simulation summarises locations of agents and other detailed information into four text files (Figure 1). At this stage, the number of households who live in sub-district *i* but have moved and stay in sub-district *j* is recorded in cell *i,j* of the population relocation matrix (see Figure 3 lower left). Data in this relocation matrix is calculated as a percentage and written into a `SocialMatrix` file (Figure 3 lower right). Finally, all four output files are used by the shelter allocation module to produce an evacuation plan.

3 Evacuation and Shelter Planning Method

3.1 Shelter Allocation Methodology

As depicted in Figure 1, the shelter allocation module takes the output of the ABS module and performs four key steps. The detail algorithm is presented in Table 1. Firstly, we use a web service from the Google Direction API to obtain three driving routes with their directions as a JASON object. The routes are extracted from the JASON object and stored in a origin/destination matrix, the OD matrix. Any unsafe roads obtained from the simulation are rejected as feasible routes.

Secondly, the travel time from all sub-districts to every shelter is calculated prior to applying the Floyd-Warshall algorithm to find the all-to-all shortest path. The calculation of travel time is based on the simple car-following evacuation model, but including the travel speed obtained from the Google Map Direction API. The RegionInfo file contains a list of vehicles that are used to calculate the length of vehicle processions for each affected region. After that, speed and distance in the OD matrix are used to calculate the overall evacuation time of all vehicle processions from each affected region to shelters. The results containing the evacuation times are stored in a new matrix called the cost matrix (C).

Table 1. Algorithm used in The Shelter Allocation Module.

Algorithm:
D: set of affected sub-districts
S: set of shelters
i,j: node index

1. **Obtain driving distance matrix** from all origin to destination (OD):
   Forall $i \in D \cup S$, and forall $j \in D \cup S$
   Obtain 3 travel routes and distance from node i to j using the Google Map Direction API
2. **Remove unsafe routes**:
   For all element in the OD matrix, remove the travel direction which contains unsafe routes
3. **Calculate all-to-all shortest path**:
   Applying the Floyd–Warshall algorithm on the OD matrix
4. **Obtain the vehicle list**:
   If a vehicle list is not provided, calculate the vehicle table based on % of households that have private cars and % of vulnerable evacuees who need an ambulance. Otherwise go to step 5.
5. **Calculate the cost matrix (C):**
   Use the vehicle list from step 4 to calculate the length of vehicle processions for all affected regions and calculate the time to evacuate all vehicles from each affected region to all shelters using the travel speed and distance in the OD matrix obtained from step 3.
6. **Combine the social matrix with the cost matrix:**
   If a social matrix is provided, concatenate all columns to the cost matrix. Otherwise go to step 7.
7. **Classify the affected regions into S centroids**:
   Pass the C matrix to the K-Means Clustering to classify the affected region into S groups
8. **Calculate the estimate evacuation demand**:
   Read the Affected Region file to obtain the % of people who decided to evacuate and calculate the evacuation demand for each region.
9. **Schedule affected regions of each centroid to the nearest shelters**:
   Allocate the affected region from the same group to the nearest shelter. If the shelter capacity is exceeded, the next nearest available shelter is selected.

3.2 Shelter Assignment based on K-Means Clustering

The third key step performs the travel-cost and social-based shelter assignment. In this step, affected regions are classified into S groups based on travel time and social connections, where S is the number of shelters. The combination of the Social matrix and the C matrix is used as the input features for the K-Means Clustering module. Ideally, all groups of an affected region will be assigned to common shelters.

To assign shelters to affected regions the number of evacuees is first estimated by applying the percentage of people that decided to go to shelters, relative to the initial population. To proceed with allocation, the system will try to assign an affected region from a group to the nearest shelter allocated to that group, *i.e.* the Our Shelter. If the Our Shelter cannot accommodate any more evacuees, the next nearest available shelter with the closest social connection will be selected, *i.e.* the Our Friend Shelter. The latter selection process is iteratively applied until an available shelter is found.

3.3 Linear Optimisation Model

The last key step in the shelter allocation is to generated alternative evacuation plans by using a linear optimisation model. To solve shelter assignment, we have implemented a linear optimisation model based on minimising travel time using the Simplex method. Parameters and decision variables in our model, are defined as follows.

S set of shelters
D set of sub-districts to be evacuated
I index of sub-district
J index of shelter
Ei the number of evacuees in sub-district i
Lj the capacity limit of shelter j
Cij the evacuation time

Decision variable:

$$x_{ij}\ is \begin{cases} x_{ij} \geq 1 \text{ and } x_{ij} \leq E_i \text{ if subdistrict i is assigned to shelter j} \\ 0 \text{ otherwise} \end{cases}$$

Objective function:

$$\min \sum_{i \in D} \sum_{j \in S} x_{ij} C_{ij} \quad (1)$$

Subject to:

$$\sum_{j \in S} x_{ij} = E_i \quad (2)$$

$$\sum_{i \in D} x_{ij} \leq L_j \quad (3)$$

In our model, the number of evacuees of sub-district i assigned to shelter j is used as a decision variable. The objective function, Equation (1), minimises total evacuation time based on travel cost to evacuate all evacuees from all sub-districts to their assigned shelters. The constraint in Equation (2) guarantees that the number of evacuees of a sub-district assigned to different shelters is equal to the total number of evacuees in that sub-district. The constrain in Equation (3), ensures that the number of evacuees from sub-districts assigned to a shelter does not exceed that shelter's capacity.

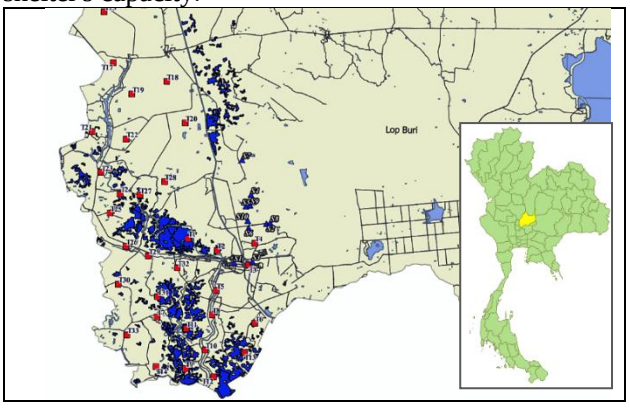


Figure 4. Screenshot showing the studied area in Lop Buri, Thailand.

4 Case Study

A case study was carried out to demonstrate the applicability of the proposed method. The case study is located in the centre of Lop Buri province, Thailand, covering 6,200 km² area. The information used in the case study is based on the Lop Buri's official evacuation and shelter plan of the year 2011. The official plan specified 33 sub-districts or Tambon (T1 – T33 shown in Table 2) with a potential flood-affected population of 149,468 people living in 8,300 households. Furthermore, there were 1,939 vulnerable people among the affected population. Ten potential public shelters were pre-defined with capacities of 500 – 3,150 people (see S1 – S10 in Table 3). Those shelters were existing public schools, hospitals, and military facilities located on high ground. Flood maps and geographical information from the Thailand Flood Monitory System provided by the Geo-Informatics and Space Technology Development Agency (GISTDA, 2016) were used to locate shelters and the affected regions (Figure 4).

The progression of the flood was simulated based on the flood map of the year 2012. To investigate advantages and drawbacks of our approach we compare the generated social-aware evacuation plans (S-plans) against their counterparts generated from the linear programming optimisation (LP-plans). We considered four scenarios (see below) that produced eight different plans: (S-A1, -A2, -B1, -B2 and LP-A1, -A2, -B1, -B2).

A1: The evacuation is performed before the inundation has progressed to the potential affected regions. Healthy evacuees use private vehicles to travel to the assigned shelters.

A2: As for A1, except that public buses are used to transport all healthy evacuees.

B1: The evacuation is started when some roads are already inundated and unsafe to use. Healthy evacuees use private vehicles to travel to the assigned shelters.

B2: As for B1, except that public buses are used to transport all healthy evacuees.

Note that ambulances were used to evacuate vulnerable evacuees in all experimental settings. The result from the ABS shows that 10 percent of the population will be evacuated. This conform with the findings in (Lei et al., 2012; Yin et al., 2014).

4.1 Shelter Utilisation

From the perspective of capacity utilisation, both the social-aware and LP- plans achieved very similar results (Figure 5) ranging from 78 – 81.5 percent. In the LP plans, we observed that some shelters were packed up to 100% to minimise the evacuation time. However, other shelters were not used at all. Thus, the social-aware plans show a slightly better (by 1%-2%) shelter utilisation.

4.2 Time to Evacuate One Household

In the generated plan, the time to evacuate one household is reported (an example of the generated plan is shown in Table 4). In all studied scenarios a household needs on average about an hour (-6 to +20 min.) to arrive at its assigned shelter (Figure 6). Although the linear optimised plans show less evacuation time per household, they have a wider range of min-max time. The most effective evacuation is observed in the S-A1 plan in which all evacuees could arrive at their assigned shelters in one hour and twenty-six minutes. The worst time is shown by the S-B1 plan as road cuts make a trip time of three hours and twenty-one minutes.

Table 2. The Demographic of Flood-Affected Regions.

ID	Sub-district	Household	Population	Vulnerable
T1	Bang Khan Mak	2,462	9,811	131
T2	Phrommat	1,275	4,563	45
T3	Tha Hin	1,655	4,629	0
T4	Tha Le Chup Sorn	2,230	9,091	116
T5	Pho Kao Ton	4,345	14,834	95
T6	Khok Lamphan	844	3,596	62
T7	Pho Tru	661	2,395	33
T8	Talung	1,051	4,234	50
T9	Si Khlong	258	993	25
T10	Ngiu Rai	1,337	5,546	103
T11	Thai Talat	880	3,477	45
T12	Kong Thanu	1,276	4,301	77
T13	Don Pho	876	4,012	23
T14	Ban Khoi	741	2,912	68
T15	Phai Yai	1,016	4,174	60
T16	Sai Huai Kaeo	530	2,287	47
T17	Mahason	744	3,069	48
T18	Sanam Chaeng	1,222	6,070	0
T19	Bang Phung	1,280	4,555	27
T20	Nong Tao	928	4,087	76
T21	Bang Kham	772	3,379	61
T22	Ban Chi	1,232	5,486	74
T23	Khao Samo Khon	1,149	4,678	110
T24	Khok Salut	500	1,912	37
T25	Bang Nga	769	3,073	31
T26	Tha Wung	594	2,075	70
T27	Muchalin	639	2,508	54
T28	Bang Li	1,082	4,519	96
T29	Bang Khu	407	1,438	24
T30	Hua Samrong	1,693	6,467	110
T31	Lat Sali	845	3,555	70
T32	Pho Talat Kaeo	1,243	4,507	71
T33	Ban Boek	1,764	5,296	0

Table 3. Shelters Information.

ID	Shelters	Capacity
S1	Khai Narai Suksa Royal Thai Army School	3,150
S2	Jirawichit Songkhram Camp	3,150
S3	The 13th Military Circle	1,050
S4	Artillery Center	2,100
S5	Artillery Division	2,100
S6	The 31st Infantry Regiment King's Guard	1,575
S7	Army Aviation	2,100
S8	Ananda Mahidol Hospital	1,050
S9	Wing 2	1,575
S10	The 11th Artillery Battalion King's Guard	525

Table 4. Example of The Generated Plan (S-A1).

From	To	Time (h:m:s)	Distance (km.)	Evacuees	Route	
T6	S1	0:21:49	11.5	361	R.3016	
T8	S1	0:18:42	9.4	425	R.3196	
T10	S1	0:23:29	15.7	557	R.3196	
T11	S1	0:24:21	14.9	349	R.5071	
T13	S1	0:18:39	17.9	402	R.3196	
T14	S1	0:30:13	22.1	291	R.311	
:	:	:	:	:	:	:
T16	S4	0:27:40	34.5	229	R.3196	
T17	S4	0:29:06	31.5	308	R.3196	
T18	S4	0:26:21	23	607	R.3196	
T31	S4	0:36:54	26.5	358	R.1	
T5	S5	0:36:30	16.3	1488	R.3196	R.1
T12	S6	0:33:21	28.9	431	R.3196	
T3	S7	0:32:35	17.4	463	R.1	
T4	S7	0:34:28	13.4	912	R.1	
T27	S7	0:26:51	22.1	253	R.3019	
T28	S7	0:27:44	14.6	453	R.3019	
T33	S7	0:44:07	36.3	530	R.3027	
T30	S8	0:35:25	28	651	R.3027	
T20	S9	0:33:18	16.7	411	R.3196	R.1
T23	S9	0:20:32	29	469	R.1	
T32	S9	0:21:22	20.8	453	R.311	R.1
T19	S10	0:18:55	30.3	458	R.3196	
T1	S10	0:35:12	19.1	985	R.3196	

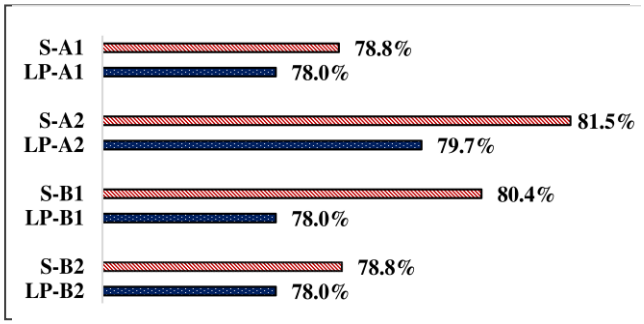


Figure 5. Average shelter utilisation of social- vs. LP-plans.

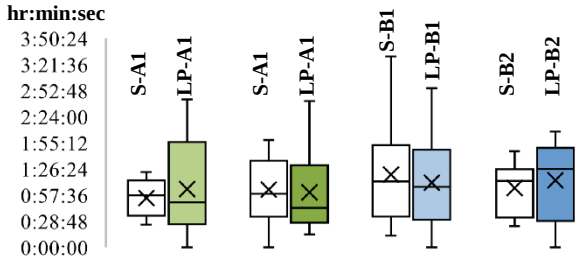


Figure 6. Time to evacuate a household to the assigned shelter.

4.3 Total Evacuation Time

The overall evacuation time is analysed from both a best case and a worst-case perspective. In the former case, termed the minimum travel time, traffic congestion was taking into account and processions of evacuees' vehicles, going to different shelters and following different routes, were scheduled in parallel. In the latter case, termed the maximum travel time, processions of evacuees' vehicles were moved in sequence to mimic heavy traffic congestion.

Travel time differences between social-aware and LP-plans are shown in Figure 7 and Figure 8. Note that positive values indicate that a social-aware plan is faster than its LP plan counterpart. Our results show that in comparison to the LP plans, all S-plans (with the exception of the minimum time of S-A1) save an average of between half an hour and up to three hours of travel time. Obviously, the travel time required for evacuation depends on the conditions of the route selected, which takes its toll in the slower minimum travel time of the S-A1 plan.

4.4 Discussion

Our approach capitalises on the detailed information on distances and available routes to produce social aware evacuation plans that seem to be more capable than simple LP plans. Note that our methodology manages to make evacuation plans even in the event of road cuts, providing that those road cuts do not cause isolation areas.

5 Conclusions and Future Work

In this paper, a GIS-enabled agent based simulation to create social-aware evacuation plans, based on estimation of household movement during flood, has been presented. The implementation of an updated version of the TU-FS simulator and the shelter assignment method have been described. A case study has been conducted to demonstrate the applicability of the proposed method. Two evacuation scenarios were considered: evacuation prior to the arrival of a flood, and evacuation during a low-level flood. The evacuation plans generated by the proposed method were compared against their linear model travel-cost optimisation counterparts. The results show that plans generated with our methodology offer a 1-2% better use of shelter capacity. The average evacuation times per household are similar to those of optimised plans with a difference between -12% to +15%. However, our evacuation plans could save nearly three hours to evacuate all regions.

Thus, the methodology presented in this paper could be an appropriate alternative to develop evacuation plans at policy level. Future work will focus on improving the social matrix model and model verification.

Acknowledgements

We thank GISTDA for the GIS data files used in this project. We thank the reviewers for their valuable comments. We thank contributors to the MASON and

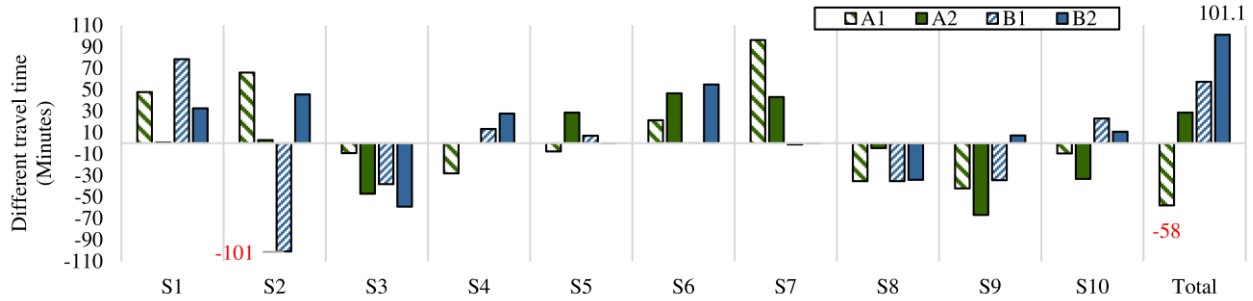


Figure 7. Difference of minimum travel time between the Social- and the LP- plans (positive value = the Social Plan is faster for x min).

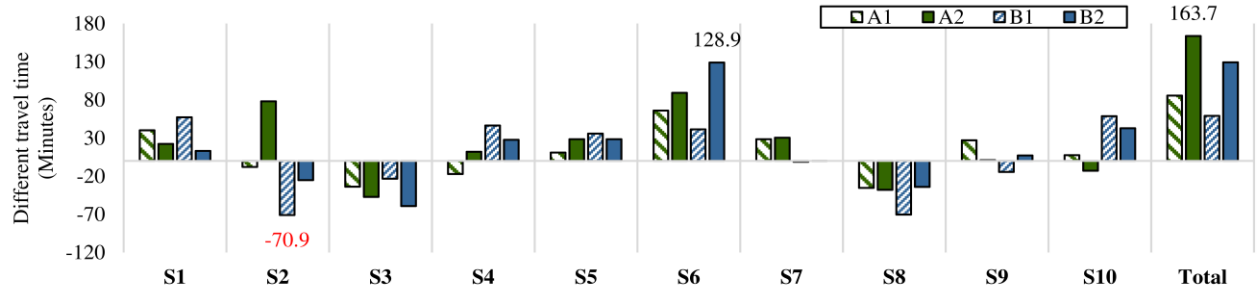


Figure 8. Difference of maximum travel time between Social- and the LP- plans (positive value = the Social Plan is faster for x minutes).

GeoMASON community website for discussion and lesson learned. We thank Professor Roland Ibbett and JC Diaz Carballo for improving the readability of this paper.

References

J. W. Bae, S. Lee, J. H. Hong, and I. C. Moon. Simulation-based analyses of an evacuation from a metropolis during a bombardment. *Simulation*, *90*(11): 1244-1267, 2014.

G. Flötteröd and G. Lämmel. Evacuation simulation with limited capacity sinks: An evolutionary approach to solve the shelter allocation and capacity assignment problem in a multi-agent evacuation simulation. In *ICEC 2010 - Proceedings of the International Conference on Evolutionary Computation*, 2010.

M. Gama, B. Santos, and M. Scaparra. A multi-period shelter location-allocation model with evacuation orders for flood disasters. *EURO Journal on Computational Optimization*, *4*(3-4): 299-323, 2016.

George Mason University. *GeoMason: GeoSpatial Support for MASON*. [accessed 2013].

GISTDA. *Thailand Flood Monitoring System*. Available via http://flood.gistda.or.th/ [accessed 2016].

Google Developer. *Google Map Direction API*. Available via https://developers.google.com/maps/documentation/-directions/ [accessed 2016].

G. Hofinger, R. Zinke, and L. Künzer. Human Factors in Evacuation Simulation, Planning, and Guidance. *Transportation Research Procedia*, *2*: 603-611, 2014.

D. Lei, W. Wenjun, and Z. Xiankun. An agent-based decision-making model in emergency evacuation management. *Journal of Convergence Information Technology*, *7*(10): 197-205, 2012.

M. B. B. Lim, H. R. Lim, M. Piantanakulchai, and F. A. Uy. A household-level flood evacuation decision model in Quezon City, Philippines. *Natural Hazards*, *80*(3): 1539-1561, 2016.

S. Luke. *Multiagent Simulation and the MASON Library*. Available via https://cs.gmu.edu/~eclab/projects/mason/.

H. S. Na and A. Banerjee. An agent-based discrete event simulation approach for modeling large-scale disaster evacuation network.In *Proceedings - Winter Simulation Conference*, 2015.

S. Vijitpornkul and W. Marurngsith. Simulating crowd movement in agent-based model of large-scale flood.In *Advanced Informatics: Concepts, Theory and Applications (ICAICTA), 2015 2nd International Conference on*, 2015.

T. Wachtendorf, M. M. Nelan, and L. Blinn-Pike. Households and Families. In *Social Vulnerability to Disasters, Second Edition*: Taylor & Francis, 2013.

W. Yin, P. Murray-Tuite, S. V. Ukkusuri, and H. Gladwin. An agent-based modeling system for travel demand simulation for hurricane evacuation. *Transportation Research Part C: Emerging Technologies*, *42*: 44-59, 2014.

A Simulation Model for the Closed-Loop Control of a Multi-Workstation Production System

Juliana Keiko Sagawa[1] Michael Freitag[2]

[1]Production Engineering Department, Federal University of São Carlos, Brazil (e-mail: juliana@dep.ufscar.br)
[2]BIBA - Bremer Institut für Produktion und Logistik, Faculty of Production Engineering, University of Bremen, Germany (e-mail: fre@biba.uni-bremen.de)

Abstract

In this paper, we propose a simulation model with a PI controller to analyze and control the dynamics of a multi-workstation production system. The formulation is based on dynamic modelling and control theory, and the model was implemented in Matlab and Simulink. Exploratory tests were carried out, and the results indicated some relationships between the values of the parameters of the controller and the values of the output variables, that is, the levels of work in process. They also showed that the proposed model has the potential of providing managerial directions on how to dynamically adjust the capacity, aiming to smooth the operation of the shop floor and to keep the work in process close to the desired levels.

Keywords: production systems, control theory, dynamics, simulation, planning and scheduling

1 Introduction

As known, the increasing computational capacity engendered a sound evolution of the operations management area, since it allowed the development of several tools to cope with large amounts of data related to the planning process and to ground the analysis of the decision makers. In this sense, the use of dynamic modeling and simulation techniques complements the static approaches for planning optimization of production systems and supply chains. Control theory is also a correlated area whose theories and tools have been applied to production and supply chain management. Some steps in these directions are enumerated in the literature review section of this paper.

Considering the aforementioned approaches, we present in this paper a simulation model based on state equations to depict the dynamics of a multi-workstation manufacturing system. A proportional-integral (PI) controller is applied to the model, and exploratory tests are carried out.

The model basically deals with the work in process (WIP) and capacity allocation variables (represented by the processing frequency of the stations), in a plant with job shop configuration. In general terms, the control of work in process is a classical concern in the operation of job shops, since it generates more predictable cycle/throughput times, which lead to better promises and fulfilment of delivery dates, a more stable coordination of the shop floor, and more flexibility to attend changes in the customer demand. These effects are highlighted in the literature related to various methodologies in production engineering, such as just-in-time, quick response manufacturing, workload control, factory physics, and others.

The results obtained with the simulation of the proposed model provided some indications of how its parameters influence the WIP levels, and demonstrated the potential of the proposed approach to depict the dynamic relations between capacity allocation, work in process and operations smoothness in production systems.

2 Literature Review

The effort of evolving from the static to the dynamic analysis of production and supply chain systems relies on system dynamics and control theory, as mentioned. In a broader sense, system dynamics may be defined as an area of knowledge that deals with the time-varying behavior of a system (Doebelin, 1998). This includes not only mechanical, electrical, fluid and thermal, but may also include biological, manufacturing, social and hybrid systems. Control theory, on its turn, has different subareas and a range of tools for the analysis and design of closed-loop systems, where the information of the outputs is fed back to the system in order to lead it to desired goals. In the literature reviews concerning the application of control theory to production and supply chain, the models are classified according to the area of application [(Ortega and Lin, 2004), (Sagawa and Nagano, 2015b)], the underlying control methodology (Sarimveis et al., 2008), the type of analysis that is carried out (i.e. robustness, stability, etc.) (Ivanov and Sokolov, 2013), or according to mixed criteria (Åström and Kumar, 2014).

In Table 1, we present some applications of control theory to production and supply chain systems, classified according to subareas of application and the methods underlying the models. Our intention here is not to present an extensive review, but rather to enumerate different possibilities and to provide few references in each category, as examples.

Other applications based on model predictive control or robust optimal control are not listed here, but can be found in (Sarimveis et al., 2008). Also, there are some alterna-

Table 1. Some Applications of control theory in production systems and supply chains

Area/type of application	*Applied methodologies and tools*	*References (examples)*
single-product production-inventory models and extensions to supply chain	classical control theory, block diagrams, transfer functions	(Towell, 1982; Zhou et al., 2006; Spiegler et al., 2016)
production-inventory models with single and multiple-machines (with or without additional constraints)	dynamic programming and optimal control	(Scarf, 1960; Boukas and Liu, 2001; Gharbi and Kenne, 2003)
multi-echelon production-inventory models using bills of material as input	input-output analysis, Laplace or z-transform, probability distributions, NPV	(Axsäter, 1976; Grubbström and Molinder, 1994; Grubbström et al., 2010)
multiple-machine and multi-product systems based on flow models	flow models, block diagrams, transfer functions, bond graphs	(Wiendahl and Breithaupt, 2000; Kim and Duffie, 2006; Sagawa and Nagano, 2015a)
supervisory/process control of continuous production systems and its integration within the hierarchical production planning	mixed integer dynamic optimization (MIDO), mixed integer nonlinear programming (MINLP)	(Monfared and Yang, 2007; Munawar and Gudi, 2004; Du et al., 2015)
production and supply chain models with autonomous control/decentralized agents	queue length estimator (QLE), pheromone, heuristic methods, RFID	(Scholz-Reiter and Freitag, 2007; Wang and Lin, 2009; Barenji et al., 2014; Schukraft et al., 2016)

tive formulations out of the control theory area, based, for instance, in queueing systems, which are out of the scope of this paper.

In the following subsection, we present the mathematical model that was adopted as basis for the simulation model proposed in this paper.

2.1 Dynamic multi-workstation model based on electrical components

A dynamic model based on the ideal properties of electrical components is proposed in (Sagawa and Nagano, 2015a) to depict a multi-workstation system that can manufacture different families of products. The model is basically composed by machines, buffers and junction elements.

The machines are compared to resistors and their processing frequency U_i correspond to $\frac{1}{R}$, where R is an ideal resistance. Similarly, the buffers are seen as ideal capacitors with capacitance C, which corresponds to their storage capacity (Ferney, 2000). The junctions are used to couple these manufacturing elements and to depict the configuration of the production flow in the system, i.e. to represent the different process routings of each product or product family (Sagawa and Nagano, 2015a). When a given machine outputs flow to m workstations downstream, it is coupled to these workstations by means of a divergent junction that imposes the conservation of flow. Similarly, the upstream flows coming from different workstations to a given workstation are merged by means of a convergent junction that conserves the total flow (Sagawa and Nagano, 2015a; Ferney, 2000). The discussed model is continuous and deals with 3 variables: the production flow f, the production volume q and the effort e. The production volume corresponds to the integral of the flow, and the effort is used as an auxiliary variable, for coupling a machine and its precedent buffer, as well for the approximation of a discrete system as a continuous system (Ferney, 2000). The basic equation of the model is derived from the well-known constitutive equations of the ideal electrical components previously mentioned, and is shown in (1). The variables and parameters of this equation were already mentioned in the text. The index i denotes a given workstation, the index s applied to the flow or effort variables denotes the output of this station, and the index e denotes its input. Eq. (2) is based on the aforementioned integral relation between the production volume q and the flow variable f, likewise the electric charge stored in a capacitor is defined as a function of the integral of the electric current. In the context of manufacturing, $\dot{q}_i(t)$ is interpreted as a rate of material storage or consumption, expressed as the difference between the input and output flow of a workstation.

$$f_{si} = U_i \left[\frac{q_i(t)}{C_i} + \min\{1, q_i(t)\} - e_{si}(t) \right] \quad (1)$$

$$\dot{q}_i(t) = f_{ei}(t) - U_i \min\{1, q_i(t)\} \quad (2)$$

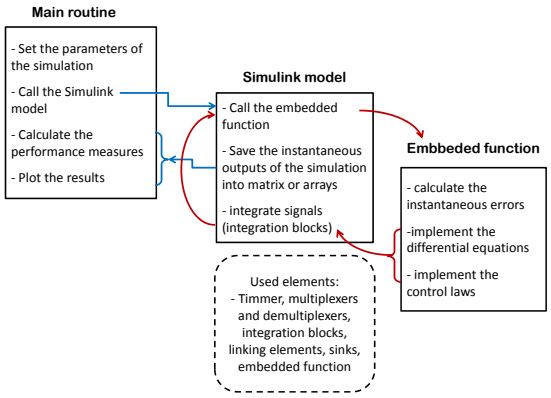


Figure 1. Schematics of the simulation model

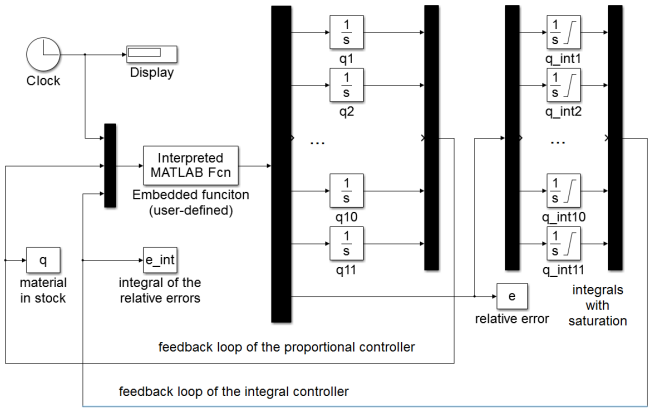


Figure 2. Schematics of the model build in Simulink®

The assumption of buffers with unlimited capacity allows simplifying (1), and the combination of (1) and (2) yields the basic state equation of a workstation, presented in (3).

$$\dot{q}_i(t) = f_{ei}(t) - U_i \min\{1, q_i(t)\} \tag{3}$$

This basic equation and the constitutive equations of the junctions were applied to a 11-workstation production system, as presented in (Sagawa and Nagano, 2015a), resulting in the state model shown in (4).

$$\begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \\ \dot{q}_3 \\ \dot{q}_4 \\ \dot{q}_5 \\ \dot{q}_6 \\ \dot{q}_7 \\ \dot{q}_8 \\ \dot{q}_9 \\ \dot{q}_{10} \\ \dot{q}_{11} \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.4300 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2508 & 0.5118 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1538 & 0.3140 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0170 & 0.0347 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0175 & 0.0020 & 0.2811 & 0.2811 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0.0087 & 0.0010 & 0.1399 & 0.1399 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0.0161 & 0.0018 & 0.2588 & 0.2588 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0.0200 & 0.0023 & 0.3204 & 0.3204 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.2407 & 0.2407 & 0.2407 & 0.2407 & -1 \end{bmatrix} \begin{bmatrix} U_1 \min\{1, q_1\} \\ U_2 \min\{1, q_2\} \\ U_3 \min\{1, q_3\} \\ U_4 \min\{1, q_4\} \\ U_5 \min\{1, q_5\} \\ U_6 \min\{1, q_6\} \\ U_7 \min\{1, q_7\} \\ U_8 \min\{1, q_8\} \\ U_9 \min\{1, q_9\} \\ U_{10} \min\{1, q_{10}\} \\ U_{11} \min\{1, q_{11}\} \end{bmatrix} + \begin{bmatrix} U_{01} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{4}$$

3 Simulation Model with a PI Controller

A simulation model based on the presented state equations was implemented using Matlab® and Simulink®. It included a proportional-integral (PI) controller, as mentioned. The structure of this model, as well as the executed instructions, are shown in Fig. 1. As it can be seen, the relevant parameters of the simulation are defined in the main routine. After that, this routine calls the Simulink model (Fig. 2), which contains the block diagram of the dynamic model with the controller. The computation of the state equations is performed by a user-defined function embedded in the Simulink® model.

After the iterations are carried out for the total simulated time, the main routine compiles the results and calculates the performance measures. For the implementation of a proportional-integral controller, integration blocks $(1/s)$ of the first level should be applied to the instantaneous material storage rates $\dot{q}_i$, while second level integrations should be applied to the relative errors in the stock levels, as it can be seen in Fig. 2.

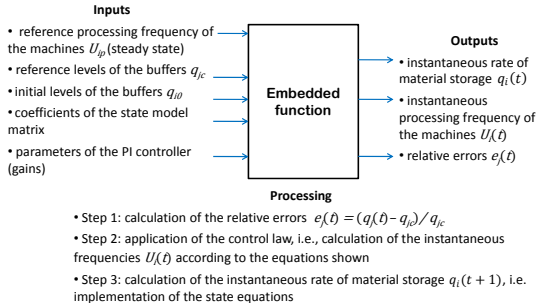


Figure 3. User-defined function implemented in the simulation model

As mentioned, the state model is implemented by means of a user-defined function. The inputs of this function are those parameters defined in the main routine, and presented in Fig. 3.

With these inputs, the function calculates, at each time t, the relative errors of the stock levels, shown in (5), and implements the control law. For an integral controller, this control law is shown in (6). With the values of the instantaneous processing frequencies of the machines U_i, resulting from the implementation of the control law, the instantaneous rates of material storage $\dot{q}_i(t+1)$ are then calculated. These rates are integrated in the integration blocks and fed back to the model

$$e'_j(t) = -e_j(t) = \frac{q_{jc} - q_j(c)}{q_{jc}} \tag{5}$$

$$U_i(t) = U_{ip}\left(1 + k_p e'_j(t) + k_i \int e'_j(t)dt\right) \tag{6}$$

where $\dot{q}_j(t)$ is the instantaneous amount of material stored in buffer j, $e'_j(t)$ is the relative error considering the actual level $\dot{q}_j(t)$ and the reference level q_{jc};U_{ip} is the reference for the processing frequency of machine i, considering the customer demand fulfillment in the steady state; k_p is the proportional gain of the controller; and k_i is the integral gain. The presented equations apply for the case where the buffer j immediately succeeds the machine i. If machine i is succeeded by more than one buffer, the minimum value of e_j is computed.

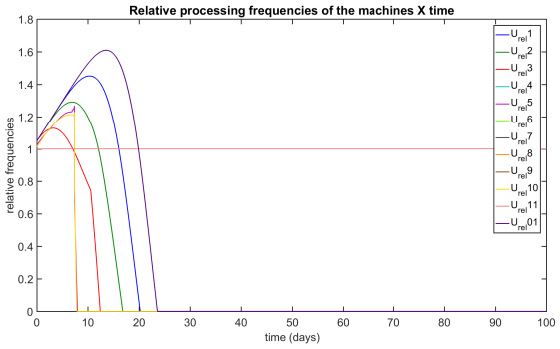


Figure 4. Processing frequencies of the machines for a test carried out with $k_p = 0.05$ and $ki = 0.05$ (whitout saturation limits).

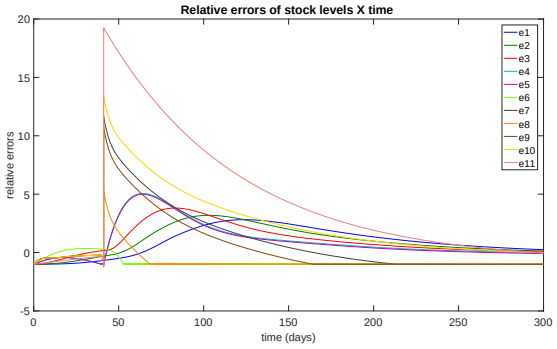


Figure 5. Evolution of the relative errors in stock levels (WIP), for saturation limits of ± 10, $k_p = 0.05$ and $k_i = 0.001$.

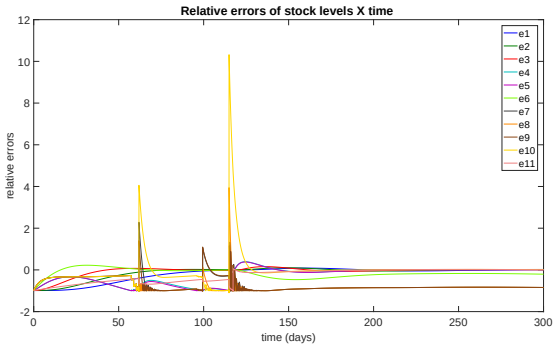


Figure 6. Evolution of the relative errors in stock levels (WIP), for saturation limits of ± 1, $k_p = 0.05$ and $k_i = 0.001$.

4 Test and Results

The proposed simulation model was applied to the 11-workstation system presented in (Sagawa and Nagano, 2015a). For an initial analysis, it was of interest to consider the warm-up of the manufacturing system and its transition to the regular operation, that is, to consider the situation where all the buffers are empty ($qi0 = 0$ for all i) and the system starts to work, aiming to attend the customer demand and to reach the desired levels of work in process. This starting condition is somewhat similar to the application of a step input, conventionally used for the study of the response in dynamic systems. In our case, however, each buffer has a different reference level, since these levels were defined as a multiple of the amount of material cumulated in each buffer when the system was simulated without control, i.e., when the open-loop system was simulated. In order to allow comparisons, we adopted a multiplication factor of 100 times, as in (Sagawa and Nagano, 2015a). As output variables of the tests, we analyzed the values of the processing frequencies of the machines $U_i(t)$ (the controlled variables); the relative processing frequencies, i.e. $(U_i(t) - U_{ip})/U_{ip}$; and the relative errors in the buffer levels (e_i) over time. This last measure indicates the variation of the work in process in the system. Depending on the selected values of the gains, the machines are led to operate with processing frequencies above the reference frequencies, in order to fulfill the buffers. In Fig. 4, the source of material works with a processing frequency that is 60% greater than the frequency that attends the customer demand in the medium term, i.e. in the steady state. This control command generates a surplus of material in the buffers. When the controller receives this information, it reduces the processing frequency of the machines. This reaction, however, is excessive, so that the machines are shutdown. The saturation of the integral controller could be a relevant parameter of influence for the control of the processing frequencies of the machines. Due to the saturation in PI controllers, the integrator may drift to undesirable values, since it tends to produce progressively larger control signs. This effect is known as windup of the integral controller (F Franklin et al., 1994; Moreno-Valenzuela, 2008). Therefore, additional tests were performed with the establishment of saturation limits for the integral controller. Some results are shown in Fig. 5-8. The values of the gains were kept constant and different saturation limits were tested.

The presented results indicate that, with narrower saturation limits, the overshoots in the WIP (represented by the relative errors) were significantly reduced. In Fig. 5, the overshoot is of 20 times the reference level and in Fig. 5, it is of 10 times. Although punctual instabilities in the control of some machines arose (Fig. 7), the operation of the system also became smoother with the estab-

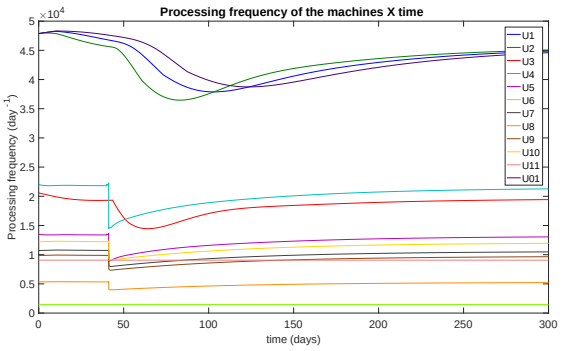


Figure 7. Evolution of the processing frequency of the machines, for saturation limits of ± 10, $k_p = 0.05$ and $k_i = 0.001$.

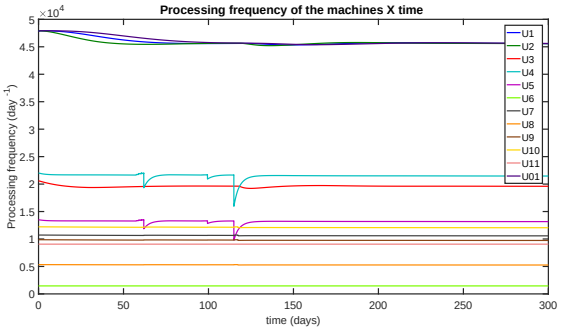


Figure 8. Evolution of the processing frequency of the machines, for saturation limits of ± 1, $k_p = 0.05$ and $k_i = 0.001$.

lishment of saturation limits.

The exploratory tests have also shown that the overshoot in the processing frequencies of the machines tend to increase with the increase of the integral gain. Hence, the results could be further improved by means of systematic experiments and the application of control tools for the optimization of these parameters (gains and saturation limits).

From the managerial perspective, the controller can lead the system to undesired operating points, depending on parameters set, because working either too far above or below the regular capacity imply higher operational costs. Moreover, there are well-known relations among WIP, cycle times and throughput rate. An increase in WIP may engender a relevant increase in the cycle time (lead time), without producing any effect in the throughput rate of the production system (Hopp and Spearman, 2001). Longer cycles times usually result in delayed jobs, rush orders, and difficulty of coordination. Based on the aforementioned reasons, it is of interest, in our model, to find the parameters that enable to reduce the overshoot of WIP and, meanwhile, to reduce the oscillations in the processing frequencies of the machines.

The variations in the processing frequencies represent, in fact, capacity additions or reductions. Thus, the proposed simulation model may provide insights on how much capacity should be increased or decreased over time in order to: 1. guarantee adequate levels of WIP, to absorb fluctuations; 2. avoid an excessive increase in the cycle/throughput times; 3. smooth operations, so that the operational costs stay low. These capacity adjustments can be implemented in various ways, i.e. overtime, subcontracting, adding/renting extra resources, or reducing the utilization of the machines.

In practice, the goal of production managers is to keep the operations stable, as much as possible, so that no extra costs are incurred (although this pursuit of stability should not unreasonably compromise the flexibility to attend customers). In MRP systems, the short term variations in the plans are usually smoothed or prevented by applying time fences to demand, planning and/or order release. This solution is efficient to keep the costs under control, but disregards the dynamics of the production systems, so that backorders and stock outs in the short term may occur. In this sense, the use of the dynamic modeling and simulation seems to be an interesting alternative or complement to other methodologies used in the production engineering area. The presented formulation is also relatively simple and easy to implement, which is an advantage in terms of use.

In order to implement it, the production system under consideration must be first modeled according to the methodology proposed in (Sagawa and Nagano, 2015a), which requires data related to the production routings, the historical demand for the end products, the product mix and the capacity of the machines (for more details, please refer to (Sagawa and Nagano, 2015a)). The mentioned methodology presents a generalization capability due to its modularity. The basic manufacturing entities, each one associated to its respective constitutive equation, may be arranged to represent different shop floor configurations, such as single machine, parallel machines, flow shop, job shop and open shop. The resulting state model will be a combination of the expressions that represent the involved entities. After this model is defined and implemented in a software for dynamic simulation of continuous systems, the adequate parameters of the controller must be defined, aiming to reduce the WIP levels and to smooth the oscillations. The generalization of the presented model requires an endeavor in this direction, since for each particular manufacturing system, a different type of controller with specific tuning could provide the best results. Thus, the control synthesis for different manufacturing systems is still an issue to be tackled.

The analysis of the results, especially in terms of the relative processing frequencies of the machines, show how much and when the capacity of each workstation should be increased or reduced, in order to achieve the desired levels of WIP. One practical limitation of the model refers to the level of aggregation of the data. Therefore, it can indicate that, for a certain period of time, a given station should work 5% above its regular capacity, but it does not give indications regarding the detailed scheduling level, i.e. indications about which specific jobs/products to process with this extra capacity, in which sequence, and so on. In other words, the model is suitable for the planning level, instead of for the detailed execution level. In order to overcome this limitation, it could be used together with discrete event simulation models, or future efforts could be undertaken towards incorporating variables that concern the scheduling level, such as set up times of the machines or processing times of individual jobs.

5 Final Remarks

In this paper, we proposed a closed-loop simulation model with a PI controller to depict the dynamics of multi-workstation production system. The model was implemented and simulated in Matlab® and Simulink® con-

sidering the warm-up of the system, when the initially empty buffers should be filled to desired levels, while the medium-term customer demand is fulfilled.

The results of exploratory tests showed that the saturation limits of the integral controller exert a relevant influence in the reduction of the work in process, but may also introduce some punctual instabilities. In future works, this parameter and the gains of the controllers could be simultaneously optimized, by means of the application of control theory tools and the execution of systematic experiments.

In terms of operations management, the proposed simulation model has the potential to give prescriptive directions about the dynamic adjustment of the capacities, in order to keep the WIP in the desired levels and the production costs relatively low, when the smoothing of capacity variations is set as a goal. In conventional production planning and control systems, based on MRP, the short term variations in production are avoided by means of the implementation of time fences to demand, planning or order release, disregarding the dynamics of the system and its ability to react to disturbances. In this sense, the presented tool can complement the existing tools for analysis and control of production systems and supply chains, allowing to take the perspective of the dynamics into consideration.

Acknowledgements

J.K.S would like to thank CNPq (Brazilian National Council for Scientific and Technological Development) for supporting this research (Grants 200648/2015-2).

References

Karl J Åström and P R Kumar. Control: A perspective. *Automatica*, 50(1):3–43, 2014.

Sven Axsäter. Coordinating Control of Production-Inventory Systems. *International Journal of Production Research*, 14 (6):669–688, 1976.

Reza Vatankhah Barenji, Ali Vatankhah Barenji, and Majid Hashemipour. A multi-agent RFID-enabled distributed control system for a flexible manufacturing shop. *The International Journal of Advanced Manufacturing Technology*, 71(9):1773–1791, Apr 2014. ISSN 1433-3015. doi:10.1007/s00170-013-5597-2.

E K Boukas and Z K Liu. Manufacturing systems with random breakdowns and deteriorating items. *Automatica*, 37: 401–408, 2001.

Ernest O Doebelin. *System dynamics: modeling, analysis, simulation, design*. Marcel Dekker, New York, 1998.

Juan Du, Jungup Park, Iiro Harjunkoski, and Michael Baldea. A time scale-bridging approach for integrating production scheduling and process control. *Computers & Chemical Engineering*, 79 (Supplement C):59 – 69, 2015. ISSN 0098-1354. doi:https://doi.org/10.1016/j.compchemeng.2015.04.026.

G F Franklin, J.D. Powell, and Abbas Emami-Naeini. *Feedback Control of Dynamic Systems*. 01 1994.

M Ferney. Modelling and Controlling product manufacturing systems using bond-graphs and state equations: continuous systems and discrete systems which can be represented by continuous models. *Production Planning & Control*, 11(1): 7–19, 2000.

Ali Gharbi and Jean Pierre Kenne. Optimal production control problem in stochastic multiple-product multiple-machine manufacturing systems. *IIE Transactions*, 35(10):941–952, 2003. doi:10.1080/07408170309342346.

R W Grubbström and A Molinder. Further theoretical considerations on the relationship between MRP, input-output analysis and multi-echelon inventory system. *International Journal of Production Economics*, 35(1-3):299-311, 1994.

R W Grubbström, M Bogataj, and L Bogataj. Optimal lotsizing within MRP Theory. *Annual Reviews in Control*, 34(1):89–100, 2010.

W Hopp and M L Spearman. *Factory Physics*. Irwin, Boston, 2001.

Dmitry Ivanov and Boris Sokolov. Control and system-theoretic identification of the supply chain dynamics domain for planning, analysis and adaptation of performance under uncertainty. *European Journal of Operational Research*, 224(2): 313–323, 2013. doi:10.1016/j.ejor.2012.08.021.

J H Kim and N A Duffie. Performance of Coupled Closed-Loop Capacity Controls in a Multi-Workstation Production System. *CIRP Annals - Manufacturing Technology*, 55(1): 449–452, 2006.

M. A. S. Monfared and J. B. Yang. Design of integrated manufacturing planning, scheduling and control systems: a new framework for automation. *The International Journal of Advanced Manufacturing Technology*, 33(5):545–559, Jun 2007. ISSN 1433-3015. doi:10.1007/s00170-006-0476-8.

Javier Moreno-Valenzuela. Experimental comparison of saturated velocity controllers for dc motors. 59:254–259, 09 2008.

S.A. Munawar and R.D. Gudi. A multi-level, control-theoretic framework for integration of planning, scheduling and rescheduling. *IFAC Proceedings Volumes*, 37(9):613 – 618, 2004. ISSN 1474-6670. doi:https://doi.org/10.1016/S1474-6670(17)31877-3. 7th IFAC Symposium on Dynamics and Control of Process Systems 2004 (DYCOPS -7), Cambridge, USA, 5-7 July, 2004.

M Ortega and L Lin. Control theory applications to the production-inventory problem: a review. *International Journal of Production Research*, 42(11):2303–2322, 2004.

J.K. Sagawa and M.S. Nagano. Modeling the dynamics of a multi-product manufacturing system: A real case application. *European Journal of Operational Research*, 244(2), 2015a. ISSN 03772217. doi:10.1016/j.ejor.2015.01.017.

Juliana Keiko Sagawa and Marcelo Seido Nagano. A Review on the Dynamic Decision Models for Manufacturing and Supply Chain. In Patricia Guarnieri, editor, *Decision Models in Engineering and Management*, pages 77–108. Springer International Publishing, Cham, 2015b. ISBN 978-3-319-11949-6. doi:10.1007/978-3-319-11949-6_5.

H Sarimveis, P Patrinos, C Tarantilis, and C T Kiranoudis. Dynamic modeling and control of supply chain systems: A review. *Computers and Operations Research*, 35(11):3530–3561, 2008.

H Scarf. The optimality of (S, s) policies in the dynamic inventory problem. *Mathematical Methods in the Social Sciences*, pages 196–202, 1960.

B. Scholz-Reiter and M. Freitag. Autonomous processes in assembly systems. *CIRP Annals*, 56(2):712 – 729, 2007. ISSN 0007-8506. doi:https://doi.org/10.1016/j.cirp.2007.10.002.

Susanne Schukraft, Sebastian Grundstein, Bernd Scholz-Reiter, and Michael Freitag. Evaluation approach for the identification of promising methods to couple central planning and autonomous control. *International Journal of Computer Integrated Manufacturing*, 29(4):438–461, 2016. doi:10.1080/0951192X.2015.1066032.

Virginia L.M. Spiegler, Mohamed M. Naim, Denis R. Towill, and Joakim Wikner. A technique to develop simplified and linearised models of complex dynamic supply chain systems. *European Journal of Operational Research*, 251(3):888 – 903, 2016. ISSN 0377-2217. doi:https://doi.org/10.1016/j.ejor.2015.12.004.

D. R. Towell. Dynamic analysis of an inventory and order based production control system. *International Journal of Production Research*, 20(6):671–687, 1982. doi:10.1080/00207548208947797.

Li-Chih Wang and Sian-Kun Lin. A multi-agent based agile manufacturing planning and control system. *Computers & Industrial Engineering*, 57(2):620 – 640, 2009. ISSN 0360-8352. doi:https://doi.org/10.1016/j.cie.2009.05.015. Challenges for Advanced Technology.

H P. Wiendahl and J W. Breithaupt. Automatic production control applying control theory. *International Journal of Production Economics*, 63(1):33–46, 2000.

L Zhou, M M Naim, O Tang, and D R Towill. Dynamic performance of a hybrid inventory system with a Kanban policy in remanufacturing process. *Omega*, 34:585–98, 2006.

Transmission of Medical Images over Multi-Core Optical Fiber using CDMA: Effect of Spatial Signature Patterns

Antoine Abche[1] Boutros Kass Hanna[2] Lena Younes[2] Nour Hijazi[2]Elie Inaty[1]Elie Karam[1]

[1]Electerical Engineering Department, University Of Balamand, Lebanon,
{antoine.abche,elie.inaty,elie.karam}@balamand.edu.lb

[2]Computer Engineering Department, University Of Balamand, Lebanon,
{boutros.kasshanna,lena.younes,nour.hijazi}@std.balamand.edu.lb

Abstract

In this work, the effect of the 2-D Optical Orthogonal Spatial Pattern Codes (OOSPC) is evaluated quantitatively for the transmission of medical images over Multi-core optical fiber using a double blind CDMA technology. The implemented method assumes that P medical practitioners or users are working simultaneously and transmitting images from one site to another. The transmitted images are encoded using a two-steps procedure: 1) coding the pixels (users) using a particular OOSPC and 2) coding the bits using time orthogonal basis functions. The encoding procedure follows the decomposition of an image into its bits to increase the transmission rate by performing a parallel transmission. Then, the encoded information from different images is combined using a multiplexer and is transmitted over the multi-core optical fiber. The transmitted information is de-multiplexed at the receiver side to identify the user that is transmitting the information and consequently, to reconstruct the original image i.e. is decoded using the same double blind Orthogonal Signatures. The performance is quantitatively evaluated using Monte-Carlo simulation techniques using different criteria, namely, the Performance Test, The Bit Error Rate, the Root Mean Square Error and the Pixel Error Rate.

Keywords: medical image transmission, CDMA, bit error rate, orthogonal spatial signature, fiber optics

1 Introduction

Over the past years, the transmission of image data over various transmission or link mediums (such as cables, local area networks, wide area networks, wireless communications and Fiber optics) has gained a big momentum [Kohli, 1989;Tsiknakis et.al., 1996]. The transmission of data, especially of a large number of images, within the same medical facility or between facilities from one site to another is of great importance in order to visualize, process and/or analyze the medical information and consequently to achieve a better diagnosis by various medical practitioners. In this context, the exchange of information will lead to the improvement of the health and the quality of the patient's life. Besides, the cost of the health care can be reduced, especially for patients who live in rural regions.

These reasons have given the impetus and the driving force to research groups and data service providers to concentrate their works in this area. Furthermore, the research has been geared toward the transmission of data using multiple access transmission technologies such as Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA) [Kohli, 1989; Tsiknakiset.al. 1996]. However, the latter techniques require very sophisticated network management and scheduling approaches [Kohli, 1989;Tsiknakiset.al., 1996]. Thus, Code Division Multiple Access (CDMA) approach has been used in image transmission to mitigate and/or to reduce such complexities [Tsiknakis et.al., 1996; Chang et. al., 1998; Abtahi et.al., 2002; Kitayana, 1994; Lisimachos et. al., 2005; Chang et. al. , 1996; Kamakura et. al., 2003; Peng et. al. , 2008; Yang et.al., 2011]. The techniques introduced in references [Tsiknakis et.al.,1996; Chang et. al., 1998; Abtahi et.al., 2002; Kitayana, 1994; Lisimachos et. al., 2005; Chang et. al. , 1996; Kamakura et. al., 2003] have used a spatial CDMA approach which in turn ensures a fast transmission. However, they have failed to preserve the pixel's intensities of the transmitted image at the receiver. The quality of the reconstructed image is of great importance in healthcare and medical applications in which a large number of images can be transmitted between various sites. That is because any distortion or error in the associated received images would lead to an inaccurate diagnosis and analysis. Consequently, wrong decisions will be made by various medical practitioners and could greatly affect the patient's life.

The CDMA-based approach has several advantages. The same range of time or the same frequency bandwidth can be occupied by several users simultaneously. Therefore, high quality images can be reconstructed using the highest spectrum efficiency in conjunction with several spatial Optical Orthogonal

Signature Patterns (OOSPs) [Yang et. al., 1998]. Consequently, the large bandwidth of the optical fiber will be an advantage in image transmission which will improve the quality of the patient's diagnostic. However, the information (data, images) should be transferred authentically and with a sufficient accuracy. Therefore, a multi access approach characterized by high speed, high spectrum efficiency and high quality data transmission rate is required for medical images' transmission. Thus, a double blind CDMA technique using a multi-core optical fiber was introduced to improve the reconstructed image's quality at the receiver end and to increase the transmission of the data rate [Abche et.al., 2011] between the various medical facilities. In this work, the effect of 2D spatial OOSPC on the latter CDMA based approach is studied.

This paper is organized as follows: the system model is presented in Section II. The method of evaluation using Monte Carlo simulation techniques is given in Section III. Also, different criteria to study its performance are introduced. The corresponding results are presented, analyzed and discussed in Section IV and a conclusion is given in Section V.

2 System Model

Figure 1 illustrates the implemented OCDMA approach to investigate the effect of 2D spatial OOSPCs on the transmission of images. It is assumed that P healthcare practitioners (users) are transmitting simultaneously medical images from one location to another over a multi-core optical fiber. Each image consists of M×N pixels and the pixel's intensity is represented by n bits. The optical encoding (temporal and spatial) and decoding of the images constitute the principal concept of the CDMA based technique [Abche et.al., 2011].

The encoding operation requires the decomposition of each image into n-binary images using the Bit Plane Decomposition (BDP) technique [Gonzalez et.al., 2008]. That is, one image is generated from the least significant bit (Bit 0) of each pixel's intensity. Another binary image consists of bit 1 of each pixel's intensity. Similar images are generated for bit 2 until bit n (i.e. the most significant bit).

The next step involves the temporal encoding of each pixel's intensity. That is, the bits are converted from serial-to-parallel for transmission purposes. Each bit is converted to an optical signal using a broadband light source (such as a light emitting diode). Then, each generated ultra-short light pulse is temporally encoded using a sequence of optical delay lines and is selected according to pre-determined basis functions. The results of each pixel's temporal encoding process (from bit 0 to bit n) are multiplexed using a passive optical coupler.

Then, the output of the multiplexer is spatially encoded using a particular 2-D spatial signature (OOSP). This procedure is crucial for user's identification by using a specific optical mask. The orthogonality between spatial signatures (OOSPs) will impose certain conditions on the spatial correlation (auto or/and Cross) in order to distinguish one 2-D spatial signature from the other signatures [G-C Yang et.al., 1998]. This double blind approach ensures the ultra high transmission rate and the accuracy needed in most medical imaging applications. The results of the spatial encoding procedure from all users are combined using a multiplexer and the corresponding information is transmitted over the multi-core fiber.

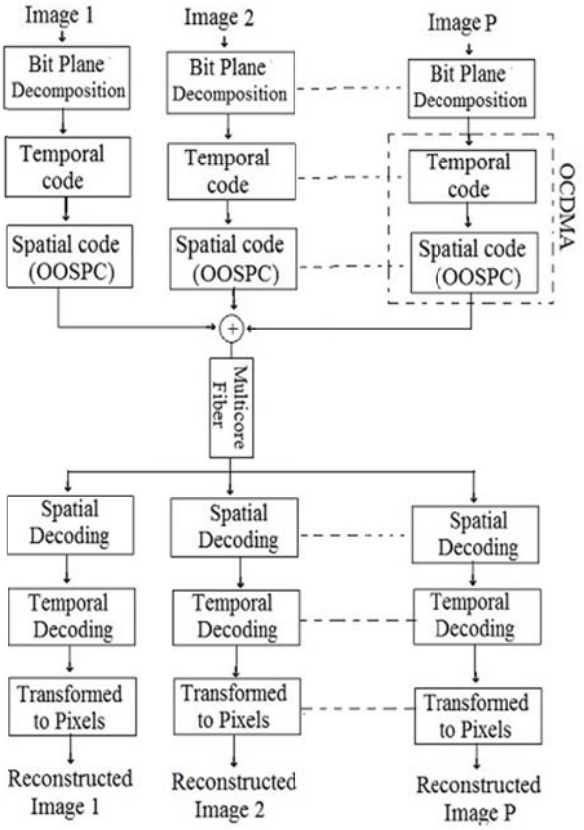


Figure 1. The implemented OCDMA approach.

At the receiver end, the signal is passively de-multiplexed using a passive optical splitter. It is fed to a set of optical decoders to identify the users, to extract the corresponding information and to reconstruct the original images. The decoding process involves the spatial decoding of the information using the same spatial mask that is used for encoding purposes. The identification of the desired user is based on the auto-correlation with the desired pixel and the cross-correlation with the interfering pixels. Then, the corresponding information is fed as inputs to n branches in which the desired signal is decoded by the same encoding temporal code. Each branch corresponds to a particular bit of the desired n-bit pixel. The decoders' outputs are converted to electrical signals using photo-detectors. Finally, the estimated

bits are converted from parallel-to-serial to form the pixels' intensities**.**

3 Method of Evaluation

The performance of the implemented OCDMA approach is quantitatively evaluated using Monte-Carlo simulation techniques. The simulated images (desired and the interfering) are generated randomly. The intensity of each pixel is represented by 8-bits and each bit is generated uniformly and randomly. That is followed by decomposing each image using the BPD approach into 8 binary images and a temporal coding is generated according to the Galois Field (GF(11)). Then, the image is encoded spatially using a 2-D OOSPC which is selected randomly from a pool of generated signatures. The OOSPCs are generated by assuming the size of the core fiber to be p by p (in this work p=7) [Yang et. al., 1998]. The pool consists of OOSPCs with a Cross Correlation (CC) and an Auto-Correlation (AC) of zero, one and/or two. Each 2-D signature is assigned to a specific image i.e. user. Having combined the information from the desired and interfering images, the results are transmitted over the optical fibers. At the receiver, the information is collected and the user is identified by performing the correlation of the desired and received patterns. Then, the coded bit stream is extracted by correlating the received data with the temporal codes that are implemented at the transmitter. Subsequently, the transmitted image is reconstructed and is evaluated quantitatively by comparing the latter with the original image because it is hampered by users' interference.

In this work, various 2-D spatial patterns are implemented to compare their effects on the transmission of medical images: the Yang &Kwong approach [Yang et. al., 1998], the Extended Hyperbolic Congruential Hop Code (EHC) [Wronskiet.al., 1996] and several versions of The Frequency-Hopped Spread Spectrum (FHSS) code [Shaar et. al., 1984.]. The performance of the presented approach is quantitatively evaluated using several similarity measures: the Bit Error Rate (BER), the pixel Error Rate (PER), the Root Mean Square Error (RMSE) and the Performance Test (PT).

3.1 Bit Error Rate (BER)

In image transmission, the BER can be defined as the percentage of bits that are transmitted in error with respect to the total number of bits. For example, a transmission with a BER of 10^{-6} means that one bit is in error when 1,000,000 bits are transmitted.

3.2 Pixel Error Rate (PER)

As it is mentioned earlier, an image consists of MxN pixels and each pixel consists of several bits. Therefore, if one bit is transmitted in error, the corresponding intensity value of the pixel will differ from its original value. In medical applications, the quality (integrity) of the received image is of great value and importance because any variation of the pixels' intensities can lead to a different appearance and consequently medical practitioners can provide an inaccurate diagnosis. Therefore, if the pixel's intensity is transmitted with an error; the reconstructed values of the received image will vary from the corresponding original intensity values. Besides, the error's severity depends on the bit's location that is transmitted in error over the multi-core optical fiber (bit 0-the least significant bit, bit 1… bit n -the most significant bit).

3.3 Root Mean Square Error (RMSE)

The RMSE is a quantitative criterion to measure the performance of a particular approach [Gonzalez et. al., 2008]. The measure is defined in terms of the difference between the original image (I(i,j)) and the reconstructed image (I_{recons}(i,j)) at the receiver end:

$$RMSE = \sqrt{\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}(I(i,j) - I_{recons}(i,j))^2} \qquad (1)$$

3.4 Performance Test (PT)

The Performance Test (PT) is a quantitative measure to evaluate the similarity of two images. It is defined in terms of two reference images: "a high quality image (I_{full}) and a low quality image (I_{worst})". The high quality image is assumed to be the original image and the latter is the objective of the transmission approach. On the other hand, the worst image is defined as to be an image where all transmitted bits are wrong. The PT is defined as [Fuderer, 1989]:

$$PT = 1 - \frac{\left\| I_{full} - I_{recons} \right\|^2}{\left\| I_{full} - I_{worst} \right\|^2} \qquad (2)$$

If the reconstructed image is similar to the original image, it is evident that the PT measure yields a value of 1. If the reconstructed is similar to the worst image, the PT has a value of 0. Therefore, the closer the reconstructed image is to the original image, the closer the PT value is to 1 and the better is the performance of the tested approach. Thus, the PT measure varies between 0 and 1 depending on the accuracy of the reconstructed image.

4 Results and Discussion

In this section, the results of the performed simulations are presented. The study is performed quantitatively using Monte-Carlo simulation Techniques. First, the principal 2-D Optical Orthogonal Spatial Pattern Codes (OOSPC) are generated by implementing the appropriate equations for each code-generation approach. Then, a pool (or pools) of 2-D spatial patterns can be constructed from the initial principal set by performing a row and/or a column shift. At this

stage, the pool consists of all the generated signatures that are correlated i.e. a pool could contain all the signatures that are characterized by an auto-correlation λa of 0 and/or 1 and a cross-correlation λc of 0 and/or 1. Consequently, the latter patterns (i.e. OOSPCs) are kept and each one will be assigned to identify a particular user (i.e. doctor) who is transmitting medical information (such as three dimensional images) over the multi-core optical fiber using CDMA approach. Thus, various pools can be generated based on the values of the auto-correlation and/or the cross correlation.

In this context, Figure 2 shows the four principal OOSPC's that are constructed using Yang and Kwong's approach and from which various pools of 2-D patterns can be generated [Yang et. al., 1998]. It is assumed that the medical images are transmitted over a 7 by 7 Multi-core Optical fiber. In each pattern, a white pixel reflects that the information is transmitted through the corresponding optical fiber. These four patterns are characterized by a correlation of zero. The generation requires the definition of two parameters and a prime number α in order to execute the corresponding equations [Yang et. al., 1998]. Similarly, the 2-D signatures of the other techniques are generated and are implemented in this work for comparison purposes.

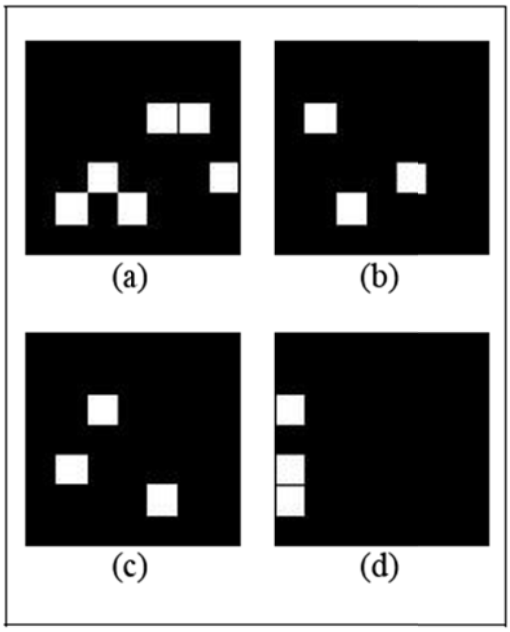


Figure 2. The principal codes generated using Yang-Kwong's approach.

Figures 3, 4, 5 and 6 show the dependence of the BER, the PER, the RMSE and the PT on the number of users (N_i) who are transmitting information simultaneously, respectively. Each figure shows the display of five plots. Each plot corresponds to a different approach: Yang-kwong (black color) EHC (blue color), FHSS-1 (red color), FHSS-2 (green color) and FHSS-3 (yellow color). The 2-D spatial signatures are divided into two pools. The first pool (Pool A) includes all the signatures that have a correlation of 0 or 1. The remaining patterns (correlation = 2) are included in the second pool (Pool B). As is already stated, the signatures are selected randomly from the first pool and then they are selected randomly from the second pool when the first pool is empty.

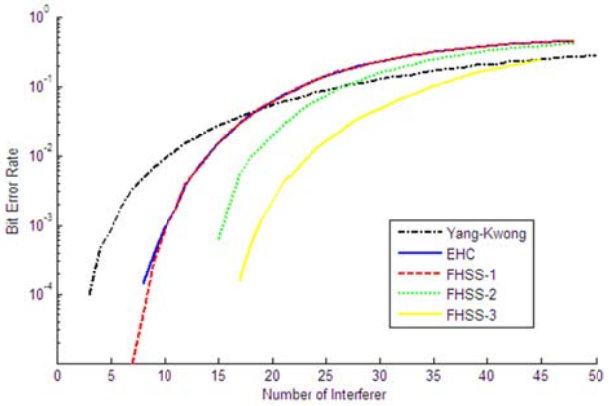


Figure 3. Dependence of BER on N_i for Various 2-D OOSPCs.

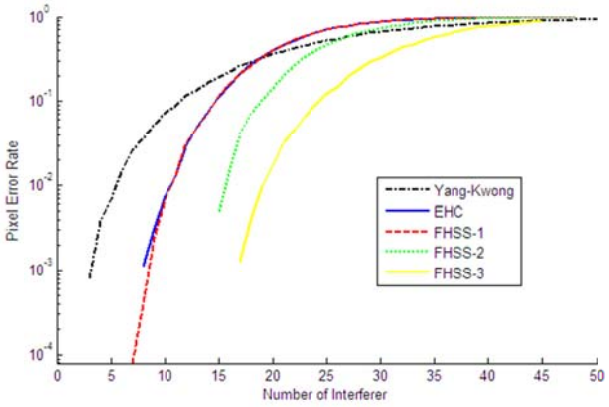


Figure 4. Dependence of the PER on N_i for Various 2-D OOSPCs.

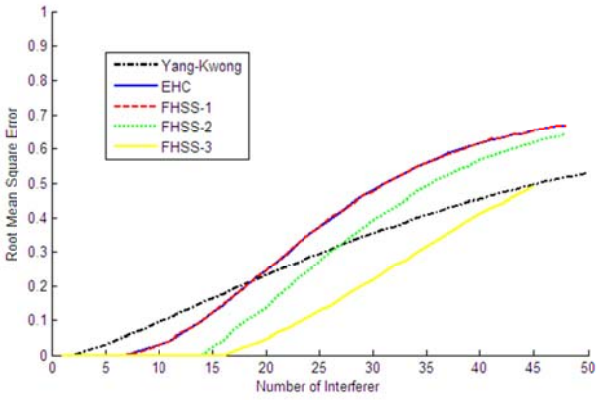


Figure 5. Dependence of the RMSE on N_i for Various 2-D OOSPCs.

The results of the Monte Carlo simulations illustrate the following:

i) The RMSE, BER and PER increase as the Number of users (N_i) transmitting over the multi-core fiber is increased. This is reflected in a higher error value. Similarly, as N_i is increased, the PT is decreased i.e. the received image is closer to the worst image. Thus, intensities of the desired image are becoming more and more different from the original intensities. This observation is associated with each 2-D OOSPC code.

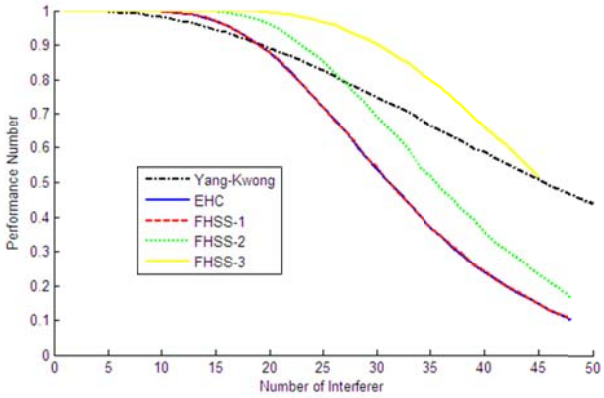


Figure 6. Dependence of the Performance Number on N_i for Various 2-D OOSPCs.

ii) The EHC and FHSS-1 based OCDMA approaches exhibit similar results for BER, PER, PT and RMSE. Besides, for a given number of users, the BER, PER, RMSE and the PT have different errors for various 2-D OOSPC codes.

iii) A transmission error is observed when 3 users are sending images simultaneously in conjunction with the desired user using the Yang-Kwong approach. On the hand, the FHSS-2 algorithm and FHSS-3 algorithm preserve the integrity of the transmitted information with 14 (FHSS - 2) and 16 interferers (FHSS - 3) , respectively (i.e. BER=0, PER=0, RMSE=0 and PT =1). Thus, it is evident that the FHSS-3 algorithm shows the best results among the five algorithms and consequently, the best encoding scheme (in this work) to send medical images over the transmission medium and to preserve the quality of these images. Also, the FHSS-2 algorithm is the second best code until 25 users are transmitting simultaneously. However, it has to be kept in mind that the FHSS-3 algorithm generates 2-D 8x8 Spatial Patterns whereas the FHSS-2 algorithm generates 2-D 7x7 signatures. This might explain the higher number of users (16 instead of 14). For completion purposes, the transmission without an error can be achieved with 7 and 8 interferers using the EHC and the FHSS-1 spatial signatures, respectively.

iv) While EHC and the FHSS-1 algorithms are the worst approaches when the number of interferers is greater than 17, the Yang-Kwong approach exhibits the worst error for Ni less than 17. That is reflected in a higher BER (Figure 3), a higher PER (Figure 4), a higher RMSE (Figure 5) and a lower performance number (Figure 6). This could be due to the fact that the corresponding 2-D OOSPCs have a higher cross-correlation.

v) The transmission error can be associated with the cross correlation at the receiver end in order to identify the desired user and the selection of the weight or threshold value. Therefore, it might be that the cross correlation differs from one family of codes to another.

As is already stated, the 2-D OOSPCs are selected from two pools and the correlation is the factor that determines the signatures in each pool. In this Monte Carlo experiment, a third pool (Pool C) is formed and it consists of all the signatures (Pool A and Pool B). Consequently, the signature is selected randomly from the pool C and is assigned to the user as outlined earlier (referred to as Scenario 1). Consequently, the corresponding results of each technique are compared with the results under the condition that the signatures are generated from Pool A and then from Pool B (referred to as Scenario 2). Figures 7 and 8 show the Bit Error Rate and the RMSE as a function of the number of interferers, respectively. The Double Blind CDMA approach is implemented with the incorporation of the FSSS-2 spatial encoding. Each figure illustrates two plots. While the first plot corresponds to Scenaio1 (blue color), the second reflects the results of Scenario 2 (green color).

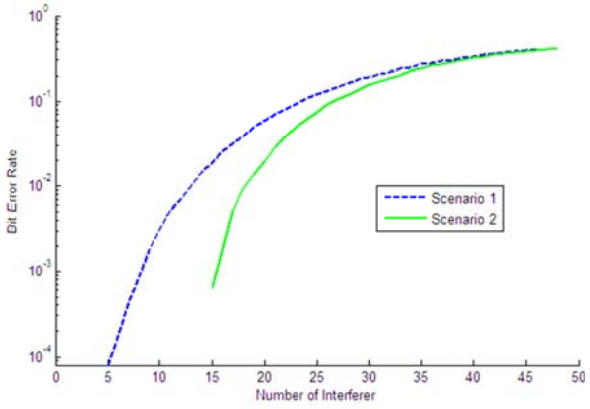


Figure 7. Dependence of BER on the selection of signatures for various number of interferer (FHSS-2 algorithm).

The results clearly show that the transmission of medical images is achieved more accurately using Scenario 2. That is, the integrity of the transmitted data is more preserved. First, as long as the number of users is less than six, the information is transmitted without an error under Scenario 1. On the other hand, fourteen users can transmit data simultaneously without any error at the receiver end. In other words, the reconstructed images will be exactly similar to the original transmitted images. This is due to the fact that the selected signatures have smaller correlation values and consequently the identification of the user is much easier. Second, for a given number of users, the BER and the RMSE values are higher for Scenario 1 than for Scenario 2. This observation remains until more than

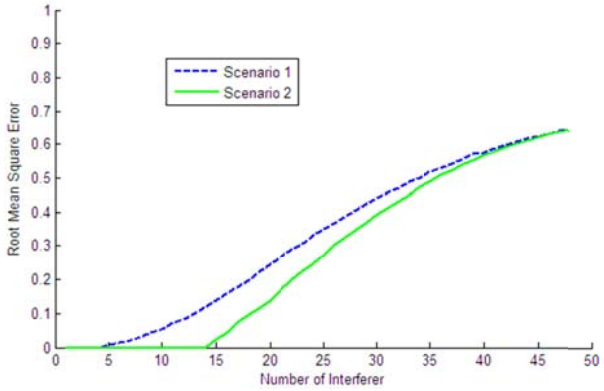


Figure 8. Dependence of RMSE on the selection of signatures for various number of interferer (FHSS-2 algorithm).

forty users are transmitting information at the same time i.e. the values of BER and RMSE become almost the same. Similar observations can be deduced for the dependence of PER and the Performance Test on the number of interferers using the FHSS-2 encoding scheme.

In the same context, Figures 9 and 10 illustrate the effect of the two scenarios on the BER and the RMSE for various numbers of users using the 2-D spatial pattern encoding FHSS-3, respectively. The corresponding results lead to the same conclusions that have been stated when the simulated images are transmitted over multi-core optical fiber and encoded using the FHSS-2 2-D spatial codes. In other words, Scenario 2 provides better results than Scenario 1. That is, the number of users working without any transmission errors is higher for the second scenario (15 vs 5) and the errors associated with scenario 2 are lower for a given number of users. Unlike the FHSS-2 based transmission approach, it can be observed that different errors exist between the two scenarios for a higher number of users. However, as it is stated earlier, this might be due to the fact that the simulated images are assumed to be transmitted over a 7 by 7 multi-core optical fiber using FHSS-2 code and over an 8 by 8 using the FHSS-3 code.

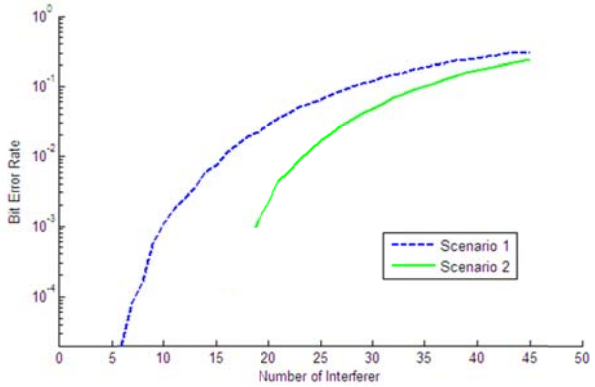


Figure 9. Dependence of BER on the selection of signatures for various number of interferer (FHSS-3 algorithm).

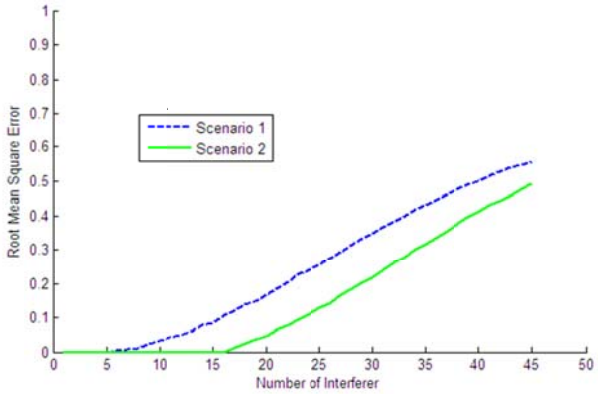


Figure 10. Dependence of RMSE on the selection of signatures for various number of interferer (FHSS-3 algorithm).

5 Conclusions

In this work, the effect of the 2-D OOSPCs on the transmission of medical images using a double blind CDMA based technology is investigated. The approach provides fast transmission rates through spatial coding and preserves the simplicity and pixels' information through the temporal coding of the pixels' intensities. It is assumed that P users are working and transmitting information (i.e. medical images) simultaneously from one site to another within the same facility or between different facilities. The main concepts of the double blind CDMA approach are the optical encoding (temporal and spatial) and the optical decoding of the images. The performance is quantitatively evaluated using Monte Carlo simulation techniques. Several measures (Bit Error Rate, Pixel Error Rate, Root Mean Square Error and Performance Test) are computed for comparison purposes. The O-CDMA transmission technique yields better results when the 2-D spatial FHSS-3 or FHSS-2 signatures are incorporated and are assigned to the medical practitioners who are transmitting the corresponding information. The latter conclusion is reflected in lower values of BER, PER and RMSE and a higher value of PT for a given number of interferers as well as in the number of users for which the error of transmission is zero. Besides, it is highly recommended to select first the signatures from a pool that are characterized by a correlation value of 0 and 1, followed by a selection from a pool characterized by a correlation of 2. Subsequently, this will lead to the transmission of images without an error with more users using any 2-D spatial encoding scheme in general, and the FHSS-2 or FHSS-3 in particular. However, it has to be mentioned that the approach based on FHSS-3 encoding scheme yields a better performance than the FHSS-2 based O-CDMA approach.

References

A. A. Shaar and P. A. Davies. A Survey of one-coincidence sequences for frequency-hopped spread-spectrum systems. *IEEE Proceedings of Communications, Radar and Signal Processing*, 131(7):719-724, 1984. DOI:10.1049/ip-f-1:19840108.

A. B. Abche, J-P Toulani, E. Inaty, and E. Karam. Medical Image Transmission Over Multicore Fiber Using Two Stages CDMA Technique. *Proceeding of the 5th International Conference on Bioinformatics and Biomedical Engineering*, pp. 1-4, May 2011, Wuhan, China. DOI: 10.1109/icbbe.2011.5780447.

C. C. Chang, H. P. Sardesai, and A. M. Weiner. Code-Division Multiple access Encoding and decoding of Femtosecond Optical Pulses over 2.5-km Fiber Link. *IEEE Photonics Technology Letters,* 10(1): 171-173, 1998. DOI:10.1109/68.651153.

G-C Yang, and W. C. Kwong, (1998). Image transmission in multicore-fiber code-division multiple access networks. *IEEE Transactions on communication*, 2(10): 285-287, 1998. DOI: 10.1109/4234.725225.

J. Kohli. Medical Imaging Applications Emerging Broadband Networks. *IEEE Communications Magazine*, 27(12):8-16, 1989. DOI:10.1109/35.41416.

Po-Rong Chang and C.-C. Chang (1996): Fiber Optic Subcarrier Multiplexed CDMA Local-Area Networks For Subband Image Transmission. *IEEE Journal on Selected Areas in Communications,* 14(9):1866-1873, 1996. DOI: 10.1109/49.545709.

K. Kamakura and K. Yashiro (2003): An embedded Transmission Scheme Using PPM Signaling with Symmetric Error Correcting Codes for Optical CDMA. *Journal of Lightwave technology*, 21(7):1601- 1611, 2003. DOI: 10.1109/JLT.2003.814387.

K. Kitayana (1994): Novel Spatial Spread Spectrum based Fiber Optic CDMA Networks For Image Transmission. *IEEE Journal On Selected Areas in Commun.,* 12(4): 762-772, 1994. DOI:10.1109/49.286683.

Lisimachos P. Kondi, D. Srinivasan, D. A. Pados, and S. N. Batalama. Layered Video Transmission over wireless Multirate DS-CDMA Links. *IEEE Transactions on Circuits and Systems for Video technology*, 15(12): 1629-1637, 2005. DOI:10.1109/TCSVT.2005.856922.

L. D. Wronski, R. Hossain, and A. Albicki . Extended Hyperbolic Congruential Frequency Hop Code: Generation and bounds for Cross-and Auto-Ambiguity function. *IEEE Transactions on Communication,* 44(3): 301- 305, 1996. DOI:10.1109/26.486324.

M. Fuderer .Ringing Artifact Reduction by an Efficient Likelihood Improvement Method. *Proceeding of SPIE, Science and Engineering of Medical Imaging*, 1137: 84-90, 1989. DOI:10.1117/12.961720.

M. Tsiknakis and D, G. Katehakis. Intelligent Image Management in a Distributed PACS and Telemedicine Environment. *IEEE Communications Magazine,* 34(7): 36- 45, 1996. DOI:10.1109/35.526886.

M. Abtahi and J. A. Salehi. Spread-Space Holographic CDMA Technique: basic Analysis and Applications. *IEEE Transactions on wireless Communications*, 1(2): 311- 321, 2002. DOI:10.1109/7693.994825.

R. C. Gonzalez and R. Wood. Digital Image Processing. , 3rd edition, Prentice Hall, 2008.

X. Peng, K-B Peng, Z. Lei, F. Chin, and C. C. Ko. Two Layer Spreading CDMA: An Improved Method for Broadband Uplink Transmission. *IEEE Transactions on Vehicular Technology*, 57(6): 3563-3577, 2008. DOI:10.1109/TVT.2008.919605.

X. Yang, and B. Vucetic. A Frequency Domain Multi-User Detector for TD-CDMA Systems. *IEEE Transactions on Communications*, 59(9): 2424-2443, 2011. DOI: 10.1109/TCOMM.2011.062111.090546.

Semantic Based Image Retrieval through Combined Classifiers of Deep Neural Network and Wavelet Decomposition of Image Signal

Nadeem Qazi B.L.Wlliam Wong

Department of Computer Science, Middlesex University, London, {n.qazi,w.wong} @mdx.ac.uk

Abstract

Semantic gap, high retrieval efficiency, and speed are important factors for content-based image retrieval system (CBIR). Recent research towards semantic gap reduction to improve the retrieval accuracy of CBIR is shifting towards machine learning methods, relevance feedback, object ontology etc. In this research study, we have put forward the idea that semantic gap can be reduced to improve the performance accuracy of image retrieval through a two-step process. It should be initiated with the identification of the semantic category of the query image in the first step, followed by retrieving of similar images from the identified semantic category in the second step. We have later demonstrated this idea through constructing a global feature vector using wavelet decomposition of color and texture information of the query image and then used feature vector to identify its semantic category. We have trained a stacked classifier consisting of deep neural network and logistic regression as base classifiers for identifying the semantic category of input image. The image retrieval process in the identified semantic category was achieved through gabor filter of the texture information of query image. This proposed algorithm has shown better precision rate of image retrieval than that of other researchers work

Keywords: image retrieval, wavelet decomposition, Gabor filter, semantic gap, stacked neural network

1 Introduction

Content-based image retrieval (CBIR) systems present a growing trend in all kind of applications including medicine, health care, internet, advertising, entertainment, remote sensing, digital libraries and crime detection. For example in medical domain, it is important to find similar images in various modalities acquired in various stages of disease progression to assist clinical decision-making process. Likewise, often during a criminal investigation, an analyst wishes to identify a digital piece of information such as unsubstantial images, tattoos, a criminal sketch generated from the details given by eyewitness or a crime scene from the huge database including both static and video images. Finding images that are perceptually similar to a query image is a challenging task in the dense database environments. There is a need for a better methodology for identifying the culprits from those images. Content-based image retrieval systems facilitate this process of image searching from large databases.

CBIR use the visual content of the query image to retrieve the best-matched image from the huge collection in the image database. The search is based on image features such as texture, shape, and color. A typical CBIR solution requires the construction of an image descriptor, which is characterized by (i) an extraction algorithm to encode image features into feature vectors; and (ii) a similarity measure to compare two images. The image retrieval system process is started by a user provided query image, followed by feature extraction of the image based on any appropriate feature selection method. A comparison of these features is then made with the features of the images in the database using some similarity measure. The matched images from the database are marked based on the value of the similarity measure, and the one having the highest value of the similarity measure is given to the user. Commonly used similarity distance measurements are Euclidean distance, Manhattan distance, Canberra distance matrix and histogram intersection distance. Researchers (Arevalillo-Herráez et al., 2008) have also suggested an algorithm to combine the similarity measurement distance based on the Bayes rule.

Previous studies on CBIR systems have focused on the global and local feature descriptor of the image. The commonly used visual descriptors are color, texture, shape and spatial relationship of the neighboring pixels in the picture. The feature extraction through a color descriptor depends on the selection of the appropriate color space. The commonly used color spaces are RGB, CIE CIE and HSV (or HSL, HSB).Color moment (Huang et al., 2010), color histogram (Sergyan, 2008) has also been used as feature descriptors in CBIR. However, a color descriptor for an image is not effective when there is a high spatial color variation. Thus, the researchers have investigated other low-level descriptors such as texture, which characterize the spatial distribution of gray levels in the pixel neighborhood.

The texture features of an image are identified by statistical, structural and spectral methods. The statistical methods for texture determination include power spectra, co-occurrence matrices, shift-invariant principal component analysis (SPCA), fractal model, and multi-resolution filtering techniques such as wavelet decomposition. However, (Selvarajah and Kodituwakku, 2011) have reported

that the first order statistical method of determining image textures are less effective in image retrieval, with an average precision rate of 0.34. Their findings showed that the second order gray level co-occurrence matrix method performed better with an average precision rate of 0.44 . The same authors have also used the coefficients of Gabor, filtered as feature vectors to retrieve the images, and reported a precision rate of 0.76. (Zheng, 2015) proposed image retrieval system named as SIMPLIcity:(Semantics-sensitive Integrate Matching for Picture Libraries). They used histogram, color layout and coefficients of wavelet transform as feature vector over 600 medical images from six categories.

The feature descriptors, whether color, texture or shape, are all low-level features and are not able to truly exemplify the high-level concept in the user's mind. This problem is known as a semantic gap in the CBIR domain and is the main hindrance in the performance of the CBIR. Image annotation, Region- Based Image Retrieval (RBIR) approaches and relevance feedback have received more attention in recent years to overcome this gap. One of the approaches used to reduce the semantic gap is image annotation. The work presented in this paper, however, is based on class based annotation, representing the image retrieval as a multiple label classification problems where each class is defined as the group of database images labeled with a single semantic label.

We have proposed in this paper, that a categorical identification based on the semantic of the query image should first be established, before the retrieval and ranking process of the similar images from the identified semantic category of the given query image. Additionally based on the fact as reported by (Cleanu et al., 2007) that human eyes use of multi-scale linear decomposition for image texture, we used multi-resolution analysis techniques to extract the feature vectors of the query image. We constructed a global feature vector using both the color and texture information of the query image through its wavelet decomposition and used this feature vector to identify the semantic category of the query image. The local texture feature of the query image was then employed through gabor filtering to create a feature vector for retrieving and ranking the similar images from the identified class. Also in order to improve the accuracy of the identification of the semantic category of the image, we have utilized the combined classifier technique consisting of deep neural network and logistic regression. This approach was later compared with the previous research studies and was found to improve the precision rate of retrieved images. This approach thus may also be helpful in the research of reducing the semantic gap, assuming that the images are labeled into classes according to the semantics of the images.

The rest of the paper is structured as follows. We present the summary of related work in section 2 followed by proposed algorithm for image retrieval in section 3. Section 4 presents feature vector extraction based on daubechies wavelet decomposition and gabor filter followed by the algorithm for combining the classifiers for image semantic identification in section 5. We compare the performance result of our proposed algorithm in section 6 and paper is concluded in the concluding segment.

2 Related Work

(Hiremath and Pujari, 2007) have combined color and texture features using wavelet-based color histograms for image retrieval from the image databases of WANG. They have used the histogram intersection distance for determining the similarity between the query image and the database image. However, an image retrieval process defined in their work uses the algorithm to retrieve the images from the whole database without any identification of image category. The performance evaluation measurement precision for image retrieval for all the categories as reported in their paper falls between 7.2 and 7.5. (Wong et al., 2007) have used support vector machines and shape-based feature extraction for image classification.

Another emerging technique in CBIR is the use of deep learning neural network. (Krizhevsky et al., 2017) has used convolution neural network consisting of five convolution layers and pooling layers having 60 million parameters and 650,000 neurons to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. (Karande and Maral, 2013) have shown relevance feedback technique using artificial neural network trained feature vectors obtained from HSV model and texture to reduce the semantic though used the cloud computing to meet the challenge of computing power. (Wan et al., 2014) have investigated towards the effective role of deep learning in reducing the semantic gap their empirical study on Caltech256 dataset has revealed that pre-trained (convolutional neural networks) CNN model on large scale dataset are able to capture high semantic information in the raw pixels and can be directly used for features extraction in CBIR tasks. They, however, concluded that features extracted by pre-trained CNN model may or may not be better than the traditional hand-crafted features.

The research work presented in this paper is however, based on prior identification of the semantic category of input query image through a deep neural network, followed by retrieving closest similar image from the identified semantic category of the image. The detailed of the proposed algorithm is presented in the next section.

3 Methodology

The proposed algorithm of the CBIR as shown in Figure 1 uses the global descriptors to extract the characteristic of the image. The algorithm works in following steps:

- Create a global features space using wavelet decomposition of HSV color space of all the images in the database.

- Identify the semantic category of the query image. This is performed through training a combined classifier consisting of a Deep Neural Network and logistic regression under supervised, using the feature vector obtained in the first step.
- In the retrieving phase of image retrieval, extract global feature vector of the query image through gabor filter.
- Retrieve and rank similar images from the identified semantic category based on the euclidean distance of gabor feature vector of the query and similar candidate images.

3.1 DataSet

For testing our proposed algorithm we have used the image data set from the Wang image database and semantically divided into 10 categories such as horse, elephants, and beaches, dinosaurs, building, food, flowers, Africa, buses, and mountains. Each division was made of approximately 100 images of the same class. We divided these image data sets into two equal training and testing data set each consisted of 500 images and used them in training and testing the classifier.

4 Semantic Identification Through Multiple Classifiers

We then created a global feature vector space of all the images in the dataset. The feature vectors were extracted using the wavelet decomposition of 2D image signal discussed below:

4.1 Global Feature Extraction through Wavelet Decomposition

Wavelet transform plays a wide role in image processing and computer graphics due to its sub-band and multiresolution decomposition ability for describing the image features and characteristics and thus one of our reasons to use it for image decomposition and feature extraction. Discrete Wavelet Transformation (DWT) uses a short life mathematical wave function (t) as its base function known as wavelets to represent a continuous time signal into different scale components. The (Daubechies, 1988) has given the following mathematical equation of the Daubechies wavelet function (t) :

$$\Psi_{rjk}(x) = 2^{0.5j}\Psi_r(2^j x - k), j, k.r \in Z$$

Where j is scale,k is a translation and r is filter.

Due to Daubechies wavelet, efficiency in separating different frequency bands and reflecting all the changes in the neighboring pixels, we chose it to extract the feature vector image signal using (HSV) color space. The Daubechies wavelet filters were convoluted with each of the images in the database using two levels of resolution, generating high- and low-frequency bands of input images. We calculated this two-dimensional wavelet image transformation by computing row by row one-dimensional wavelet transformation in a horizontal direction, and then a column by column one-dimensional wavelet transformation in a vertical direction as shown in Figure 2. This produced the first level of decomposition. For the second level decomposition, we used this same process,however, using the low-level frequency component obtained in the first level decomposition. This finally yielded two levels of high and low-level frequency components generating four sub-images, which are labeled as LL, LH, HL and HH in the Figure 2, where

- Sub-image LL1 and LL2 represent the horizontal and vertical low-frequency part of the image at level 1 and 2 respectively and are recognized as an approximation.
- Sub-image HH1and HH2 represent the horizontal and vertical high-frequency part of the image at level 1 and 2 respectively and are called diagonal.
- Sub-image LH1 and LH2 represent the horizontal low and vertical high-frequency components at level 1 and 2 respectively and are known as horizontal.
- Sub-image HL1 and HL2 represent the horizontal high and vertical low-frequency components at level 1 and 2 respectively and are called vertical.

This two-level wavelet decomposition process generated approximation coefficients along with the detail coefficients in horizontal, vertical and diagonal direction at

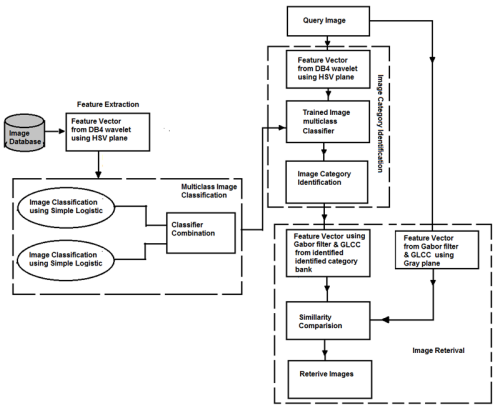


Figure 1. Proposed algorithm of the CBIR.

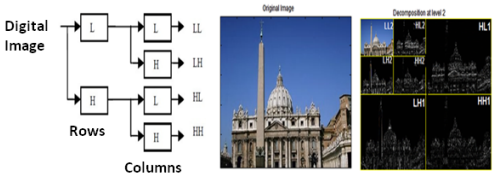


Figure 2. Wavelet Decomposition of 2D-image signal.

each level. We then constructed an image feature vector consisting of 33 elements for each of the images in the database. Each image feature vector consisted of standard deviation, mean, skewness and kurtosis of the histogram of the detail coefficients and energy vector of both the approximation and detail coefficients obtained in the two-level decomposition of the input signal. This image feature vector was later used to identify the semantic class of the query image.

4.2 Deep Learning of Neural Network

We then used the extracted global feature vector to train classifiers for identification of the semantic category of the query image. To achieve this we defined an image domain space consisting of N samples of images, each represented by a global feature vector q (f1, f2, f3....f33) to be used as input feature vector to the classifier. The goal was set to assign every input query image a semantic class label Ci from the class labels space C (C1,..C10) using these trained classifiers. We employed the deep learning approach in training the classifier to identify the semantic category of the query image. Two classifiers were explored i.e. deep neural network and logistic regression. Deep learning refers to a class of machine learning techniques that employ deep architecture, unlike their shallow counterpart processing information through multiple stages of transformation and representation. Deep learning neural network architectures are different from "normal" neural networks because they have more hidden layers.

A Deep Neural Network consisting of one input layer, three hidden layers, and one output layer each having sigmoid activation function was constructed. The choice of the activation function was made following the research work of (Shenouda, 2006). They have performed a quantitative comparison of the four most commonly used activation functions, including the Gaussian RBF network, over ten real different datasets to show that the sigmoid activation function is a substantially better activation than others. Back propagation training algorithm was employed, to 10 fold classification. Feature vectors generated through the process of Daubechies wavelet decomposition were used as external inputs to five layers deep learning neural networks during the training phase. Following equation was used to calculate network output after each layer of the selected neural architecture.

$$a^{(i+1)} = S^{(i+1)}(W^{(i+1)}a^i + b^{(i+1)})$$

Where i=1,2,3,4
$a_{,}$, is the output from i^th neural network layer
$S^{(i+1)}$, is the sigmoid function
$W^{(i+1)}$ and $b^{(i+1)}$ are neuron weights The input a^0, consisting of feature vector $f_{1,1}$, to $f_{n,33}$, to the input layer is then given as:

$$a^0 = q = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{d1} & x_{d2} & x_{d3} & \dots & x_{dn} \end{bmatrix}$$

The network outputs generated by the output layer of the network are given by

$$a^4 = C = \begin{bmatrix} C_1 & C_2 & C_3 & \dots & C_k \end{bmatrix}$$

Where n is the number of observations and the training set in the form $(q_1\ C_1), (q_1\ C_1), (q_n\ C_k)$ were presented to the feed forward neural network during training.

In addition to these, we also chose logistic regression as of our second classifier. However, in this case, we estimated a probability of a given instance Cj belonging to semantic class space C=(C1,..C10). The probability for class j excluding the last class for multi-class problems was determined by

$$p_j(f_j) = \frac{e^{f_i\theta_j}}{\sum_{j=1}^{k-1}(1+e^{f_i\theta_j})}$$

The probability of the last class was calculated using the following equation

$$1 - \sum_{j=1}^{k-1} p_j(f_j) = \frac{e^{f_i\theta_j}}{\sum_{j=1}^{k-1}(1+e^{f_i\theta_j})}$$

Where k is the number of classes, n is the total number of observations, f_i is the input feature vector and $theta$is the parameter matrix, which is calculated using the Quasi-Newton Method.

The two classifiers were then fused to improve the classification accuracy. This approach of classification through fusion is increasingly embraced by researchers in recent years. (Baskaran et al., 2004) have combined weighted multiple classifiers consisting of naive Bayes, artificial neural networks, fuzzy C-mean classifier and variants of distance classifiers for the remote sensing image classification. (Qazi and Raza, 2012) has suggested using combined classifiers for the better classification of network intrusion minor classes.

We used the stacking method to combine the trained classifiers. The stacking of classifiers algorithm is relatively a new approach in classifier combination and consists of classifiers at two levels i.e. base classifiers and Meta classifier or arbiter. The Meta classifier selects the best classifier among several base classifiers. We used linear regression as the training algorithm for the Meta learner to stack the two base classifiers i.e. deep Learning Neural Network and logistic regression. The training dataset was then used to train this fused classifier for identification of semantic image category.

5 Image Retrieval

The retrieval phase of the similar images from the matched semantic category uses textural features of the input query image. We constructed texture feature vector of the images using the localized feature of the image through Gabor filter due to its wavelet nature capturing energy at a specific frequency and a specific direction. The mathematical representation of a 2-D Gabor filter (Gb), having wavelength represented by λ, orientation angle by *theta*, and standard deviation along x and y by $\sigma_x$and $\sigma_y$respectively, may be given by the following relation:

$$\text{Gb(x,y)}= Gs(x,y)e^{j\lambda(xcos\theta+ysin\theta)}$$

Where Gs(x,y) is the Gaussian function given by

$$Gs(x,y) = \frac{e^{-0.5((x/\sigma_x)^2+(y/\sigma_y)^2)}}{\sqrt{2\pi\sigma_x\sigma_y}}$$

We generated a filter bank consisting of 36 filters by varying the wavelength λ from 2.5 to 3.0 with an increment of 0.1, and the orientation angle $\theta$from$\pi, 3/2\pi, 11/6\pi, 25/12\pi, 187/60\pi, and\ 441/180\pi$. Each of the filters in the filter bank was then convoluted with the input image finally the feature vector consisting of 144 elements was constructed by calculating the contrast, homogeneity, correlation and energy of the GLCM (Gray Level Co-occurrence Matrix) for each filtered image in the filter bank.

6 Performance Analysis

In order to assess the performance of the proposed algorithm, we used the test data set consisting of 250 images approximately 25 from each category. For each tested query image two feature vector was extracted , wavelet based feature vector was used to identify the semantic category and then another Gabor filter-based feature vector of the query image was used to retrieve the smaller images from the identified category using euclidean distance.

The precision and recall rate of the classification obtained by the combined classifier of deep neural network and logistic regression for the testing dataset is shown in Table 1. However, Figure 3 shows precision and recall rate

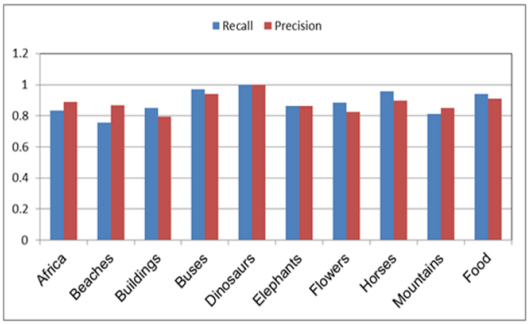


Figure 3. Classification accuracy of the combined classifier.

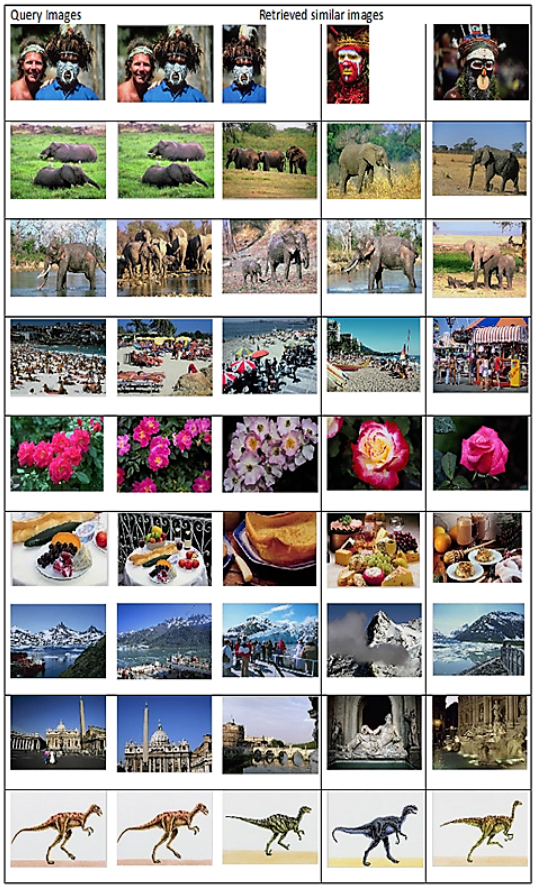


Figure 4. Result of visual queries through the proposed algorithm.

for all the semantic categories used in the test data. Some of the visual queries and retrieved similar images are presented in Figure 4. The average retrieval rate was found to be less than 1 minute approx 40 to 50 seconds for retrieving four similar images of the query image. A precision rate of all the retrieved images using a test data set has been calculated and shown in Table 2 in comparison with the algorithms of other researchers. It can be seen that our algorithm has performed slightly better it is because that the identifying the right category of query image increases the probability of the retrieving similar image.

Table 1. Trained classifiers accuracy.

Instances	ANN	Logistics	Stacked Classifiers
Correctly Classified	83.5%	85.9%	87.0%
Incorrectly Classified	16.5%	14.1%	13.0%

7 Conclusions

It was noted in this research that classification of the images in the semantic based categories classes may be help-

Table 2. Performance Comparison With Other Algorithms.

Category	Proposed algorithm %	Ahmed J.Afifi Wasam Ashour %	Manimala Singha et al %	Chen-Horng Lin et al %	M Babu et al %
Africa	0.74	0.71	0.65	0.68	0.56
Beaches	0.90	0.85	0.62	0.54	0.53
Buildings	0.86	0.83	0.71	0.56	0.6
Buses	0.82	0.85	0.92	0.89	0.89
Dinosaurs	0.99	0.99	0.97	0.99	0.98
Elephants	0.76	0.71	0.86	0.66	0.57
Flowers	0.94	0.93	0.76	0.89	0.89
Horses	0.89	0.57	0.87	0.80	0.78
Mountains	0.86	0.42	0.49	0.52	0.51
Food	0.85	0.97	0.77	0.73	0.69
Average	0.86	0.78	0.76	0.72	0.70

ful in reducing the semantic gap. It also improves the retrieval efficiency because after the related semantic class of input query image is identified then retrieval of similar images is performed within the group of more related images in the same class. We aim to apply the proposed technique in developing a robust image and video search engine that could assist the analyst in retrieving photographs, images of the criminals or crime scenes from huge criminal database such as VALCRI database.

Acknowledgment

The research leading to the results reported here has received funding from the European Union Seventh Framework Program through Project VALCRI, European Commission Grant Agreement NÂř FP7-IP-608142, awarded to Middlesex University and partners

References

Miguel Arevalillo-Herráez, Juan Domingo, and Francesc J Ferri. Combining similarity measures in content-based image retrieval. *Pattern Recognition Letters*, 29(16):2174–2181, 2008.

R Baskaran, M Deivamani, and A Kannan. A multi agent approach for texture based classification and retrieval (matbcr) using binary decision tree. *International Journal of Computing & Information Sciences*, 2(1):13, 2004.

Ctlin Cleanu, De-Shuang Huang, Vasile Gui, Virgil Tiponu, and Valentin Maranescu. Interest operator versus gabor filtering for facial imagery classification. *Pattern Recogn. Lett.*, 28(8):950–956, June 2007. ISSN 0167-8655. doi:10.1016/j.patrec.2006.12.013. URL http://dx.doi.org/10.1016/j.patrec.2006.12.013.

P. S. Hiremath and J. Pujari. Content based image retrieval using color, texture and shape features. In *15th International Conference on Advanced Computing and Communications (ADCOM 2007)*, pages 780–784, Dec 2007. doi:10.1109/ADCOM.2007.21.

Z. C. Huang, P. P. K. Chan, W. W. Y. Ng, and D. S. Yeung. Content-based image retrieval using color moment and gabor texture feature. In *2010 International Conference on Machine Learning and Cybernetics*, volume 2, pages 719–724, July 2010. doi:10.1109/ICMLC.2010.5580566.

S. J. Karande and V. Maral. Semantic content based image retrieval technique using cloud computing. In *2013 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–4, Dec 2013. doi:10.1109/ICCIC.2013.6724277.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017. ISSN 0001-0782. doi:10.1145/3065386. URL http://doi.acm.org/10.1145/3065386.

N. Qazi and K. Raza. Effect of feature selection, smote and under sampling on class imbalance classification. In *2012 UKSim 14th International Conference on Computer Modelling and Simulation*, pages 145–150, March 2012. doi:10.1109/UKSim.2012.116.

S Selvarajah and SR Kodituwakku. Analysis and comparison of texture features for content based image retrieval. 2011.

S. Sergyan. Color histogram features based image classification in content-based image retrieval systems. In *2008 6th International Symposium on Applied Machine Intelligence and Informatics*, pages 221–224, Jan 2008.

Emad A. M. Andrews Shenouda. A quantitative comparison of different MLP activation functions in classification. In *Proceedings of the Third International Conference on Advances in Neural Networks - Volume Part I*, ISNN'06, pages 849–857, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-34439-X, 978-3-540-34439-1. doi:10.1007/11759966_125. URL http://dx.doi.org/10.1007/11759966_125.

Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retreval: A comprehensive study. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, pages 157–166, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3063-3. doi:10.1145/2647868.2654948. URL http://doi.acm.org/10.1145/2647868.2654948.

Wai-Tak Wong, Frank Y. Shih, and Jung Liu. Shape-based image retrieval using support vector machines, Fourier descriptors and self-organizing maps. *Information Sciences*, 177(8):1878 – 1891, 2007. ISSN 0020-0255. doi:https://doi.org/10.1016/j.ins.2006.10.008. http://www.sciencedirect.com/science/article/pii/S0020025506003227.

K. Zheng. Content-based image retrieval for medical image. In *2015 11th International Conference on Computational Intelligence and Security (CIS)*, pages 219–222, Dec 2015. doi:10.1109/CIS.2015.61.

A Method for Modelling and Simulation the Changes Trend of Emotions in Human Speech

Reza Ashrafidoost[1] Saeed Setayeshi[2]
[1]Department of Computer Science IAU, Science and Research University, Iran, r.ashrafidoost@srbiau.ac.ir
[2]Amirkabir University of Technology, Iran, setayesh@aut.ac.ir

Abstract
One of the fastest and richest methods, which represents emotional profile of human beings is speech. It also conveys the mental and perceptual concepts between humans. In this paper we have addressed the recognition of emotional characteristics of speech signal and propose a method to model the emotional changes of the utterance during the speech by using a statistical learning method. In this procedure of speech recognition, the internal feelings of the individual speaker are processed, and then classified during the speech. And so on, the system classifies emotions of the utterance in six standard classes including, anger, boredom, fear, disgust, neutral and sadness. For that reason, we call the standard and widely used speech database, EmoDB for training phase of proposed system. When pre-processing tasks done, speech patterns and features are extracted by MFCC method, and then we apply a classification approach based on statistical learning classifier to simulate changes trend of emotional states. Empirical experimentation indicates that we have achieved 85.54% of average accuracy rate and the score 2.5 of standard deviation in emotion recognition.

Keywords: emotional speech modelling, speech recognition, human-computer interaction (HCI), gaussian mixture model (GMM), Mel frequency cepstral coefficient

1 Introduction
The manners of speaking have eminent role in human communications, which are the natural methods to express the emotion and feeling in conversation. Equally important, the tone of voice is a method to express the state of emotion of the speaker. Once, an utterance expresses the word with an emotion that makes his tone of speech change, the meaning of the word is accomplished. Up to date, Emotion recognition of speech is one of the challenging fields in modeling systems which are based on human computer user interface. These systems could simulate the feelings including uttered speech, if equipped with intelligent emotional recognition techniques and algorithms (Cowie et al., 2001). Using this kind of systems would outline the attributes of uttered speech including psychological and cognitive background, and the emotions of speaker. This approach provides the possibility for intelligent or adaptive system designers to design machines, which make suitable automatic reactions in accordance with natural human needs at different situations. Evidently, one of the major areas has attracted loads of attentions to these systems is automatic emotion recognition (AER) of human speech.

Scientists, who have been working on voice and speech technology for the past four decades, have now good understandings of the voice analysis, human speech modeling and speech processing-based systems; this leads them to develop various practical applications in this field. With regard to the capabilities which provided by speech signal analysis, researchers in the field of artificial intelligence, robotics and human-computer interaction (HCI) could design machines which would be useful to develop tools and systems, which are related to human natural behavior. Some of these systems would be similar to responsive and adaptive systems, speech production, speech simulation, evaluations systems, security and surveillance systems, speaker recognition systems, human-robot interactive systems (HRI), and generally the environments which are equipped with smart workplace systems. To achieve this purpose, it is needed to automate data collection from users to get optimal performance of these systems and could perform services real-time and compatible with user's needs.

In this paper, we propose an approach which could acquire the attributes of the utterance and his emotional changes by studying the patterns of speech signal. This information is used to recognize the conceptual characteristics of speech which wrapped in the voice of humans by an intelligent machine. In this study we apply speech corpus of utterances from EmoDB as input, and then processing of speech signal recognition begins, some pre-processing tasks are performed on raw speech signal, and then desired features are extracted by Mel Frequency Cepstral Coefficient (MFCC) method. Then, the attributes of each uttered word of speech is elicited

separately by the Learning Gaussian Mixture Model (LGMM), as the innovative classification approach. These emotional states, which stand for the emotional state of speaker for each word during speech, are labeled and arranged side by side in according to speech stream. To end with, we could delineate the trend of emotional states of speaker during speech or conversation.

By using the proposed approach, the emotional states of utterance are classified in six standard emotional classes. These classes include, anger, boredom, fear, disgust, neutral and sadness. Despite the context of expressed talk during the speech, the system could detect and track the trend of credible internal emotional states of utterance. Emotional speech recognition using this method of classification provides the precise results and high accuracy in emotion recognition and its changes trend. The most prominent goal of this article is to propose an approach based on an innovative learning method of Gaussian Mixture Model in emotional recognition of speech to extract internal emotions and feelings of the speaker. This, is performed by processing the speech signal, and then represent changes trend of emotional states. This approach of speech patterns processing, could be used in intelligent systems which closely interact with users to predict the emotional states of them. The systems which equipped with this capability could be used in the fields like medical, educational, surveillance systems or intelligent work places.

The reminder of the paper is structured as follows, Section 2, delineates some of recent related works in this field and their specifications. Section 3, illustrates the overall view of methods and concepts using in this article; including Emotion recognition; feature extraction method, MFCC; and the Gaussian Mixture Model. Section 4 introduces the database we have used during the test and train phases (Burkhardt et al.,2005). In Section 5, the computational architecture of proposed approach is presented and described the structure of method. Section 6 provides tentative results and performance measurements, and finally Section 7 draws conclusion remarks.

2 Related Works

Recognition of emotion in speech and tracking its changes trend to disclose internal feelings of speaker is the current topic in the field of artificial intelligence, signal processing, and human-computer interaction in the recent years. To the best of our knowledge, some researchers have focused specifically on localizing emotion transition wrapped in speech. Most of them focused on acoustical features of the speech signal. For example, using troughs and peaks in the profile of fundamental frequency; intensity and boundaries of pauses; and energy of signal were the popular clues to design emotion classifiers. So, we are focusing on some of recent outstanding researches in this field and briefly investigate their specifications.

Anguera et al. (2011) proposed an approach to detect speaker change, which using two consecutive fixed-lengths windows, modeling each by Gaussian Mixture Model and distance-based methods, such as Generalized Likelihood Ratio (GLR), Kullback-Leibler (KL) divergence, and Cross Log Likelihood Ratio (CLLR) have been investigated. In 2013, another study performed by C.N. van der Wal and W. Kowalczyk who proposed a system to measure changes in the emotional states of the utterance automatically by analyzing voice of speaker. They represented the obtained results by visualizing them in 2-D space. In this study the Random Forest algorithm was applied for classification and regression problems. Their results show some improvements in performance and error reduction in compare with similar studies which focused on predicting changes of intensity measured by Mean Square Error (MSE). They also claimed that the proposed system performs to classify negative emotions and provides better performance.

Besides, in the other studies for extracting emotion from speech, a number of useful methods like SVM (Fergani et al., 2008), Variational Bayes free energy (Valente, 2005) and factor analysis (Kenny et al.,2010) have used. However, it seems that these methods require large databases for testing and training phases to be effective.

3 Preliminaries

3.1 Emotion Recognition

Different moods and emotional feelings reflected in the voice of speakers are represented by the special patterns of acoustical features in speech signals. This means that the worthwhile information wrapped with emotional states of the utterance is encoded in acoustical speech signal of the voice of speaker. This information would be decoded and then embedded emotions disclosed and could be perceived and feel once receiving by audiences. Therefore, the first step to design the automatic emotional recognition systems is to find out how to encode the emotional states which expresses by the speaker in the speech. This work is done by extracting the most discriminator features from speech samples in training phase. Then the classification

method resolves this issue and decodes the data in order to recognize the class of particular emotional state (Yang and Lugger, 2009).

3.2 Feature Extraction (MFCC)

Cepstrum coefficients of Mel frequency is the representation of the speech signals which extracts the non-linear frequency components of the human auditory system. This method converts linear spectrum of speech signal to non-linear frequency scale which is called "Mel". At the first stage of our proposed method, pre-processing tasks are performed on the raw speech input signal using windowing techniques (Kowalczyk and van der Wal, 2013). The windowing is done after providing Discrete Fourier Transform (DFT) of each frame to obtain the spectrum scale of speech signal (Motamed, 2014). Then, frequency wrapping is used to convert spectrum of speech to Mel scale where the triangle filter bank at uniform space is achieved (Rahul et al., 2015). These filters multiplied by the size of spectra and eventually obtained MFCCs. In this paper 20 filter banks and 12-MFCC are used for feature extraction. Mel-scale frequency conversion equation is determined in

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \quad (1)$$

and the transpose equation of Mel frequency transformation is showed in

$$M^{-1}(f) = 700 \left(\exp\left(\frac{m}{1125}\right) - 1\right) \quad (2)$$

3.3 Gaussian Mixture Model (GMM)

In statistical sciences, the mixture model is considered as a probabilistic model which is used to represent existence of the subsets of classes which belong to the larger population. A Bayesian model like GMM is one of the special cases of these statistical models. GMM modeling technique is straightforward but so efficient. Therefore, these capabilities are significant due to its ability of forming soft approximations and curved shapes of any form of distribution in random data. This model is used as a successful model in different systems, especially in the field of speech recognition and speaker identification systems.

Accordingly, the Gaussian Mixture Modeling first invented by N. Day and later by J. Wolfe at the late 60's (Wolfe, 1970) known as Expectation-Maximization (EM) algorithm (Ververidis and Kotropoulos, 2005). Hence, the main reason of using this model in the wide range of intelligent systems is the ability of this technique to model the data classes or the distribution form of acoustical observations of the speaker (Alaie et al., 2015). According to

$$F(x|\lambda_k) = \sum_{i=1}^{K} c_i f_i(x) = \sum_{i=1}^{K} c_i \mathcal{N}(x|\Phi_i) = \sum_{i=1}^{K} c_i \mathcal{N}(x|\mu_i, \Sigma_i) \quad (3)$$

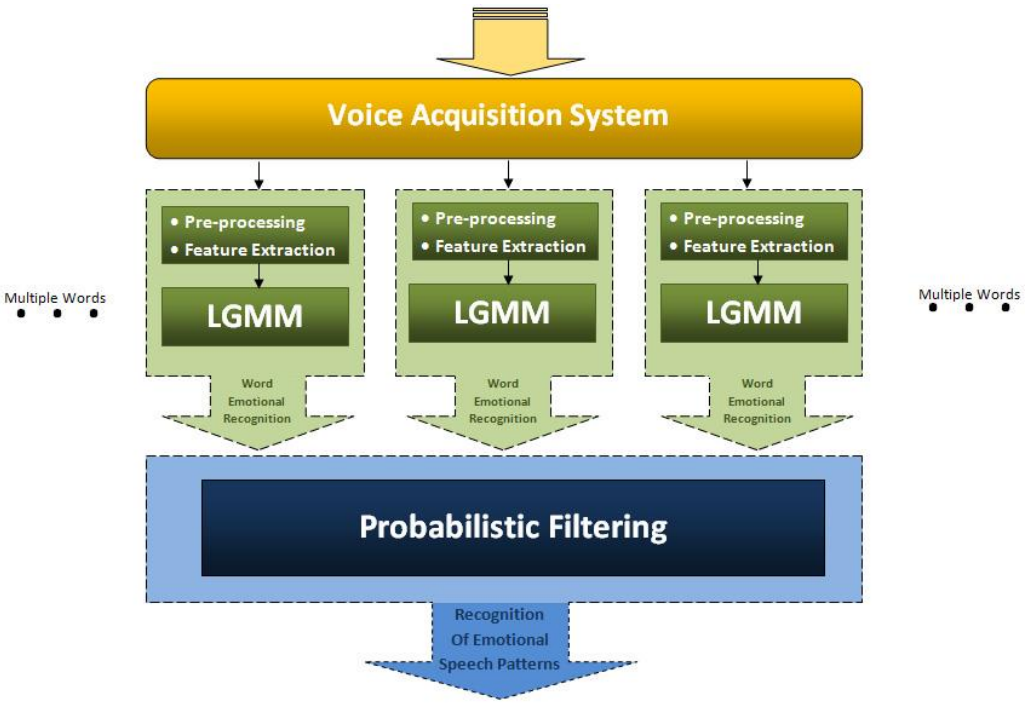


Figure 1. Block diagram of emotional speech recognition routine.

in the GMM likelihood function which has been used for D-dimensional feature vector, x is a weighted sum of K multivariate Gaussian components, $f_i(x)$, is D×1 for each mean vector (μ_i) and D×D covariance matrix (Σ_i).

In (3), λ_k represents parameters of GMM and include K components in order to the restricted states, in which the combined weights should be satisfied by the following two conditions; $c_i \geq 0$ for i=1,…,K and $\sum_{i=1}^{K} c_i = 1$. i-th component could be written as

$$f_i(x) = \mathcal{N}(x|\Phi_i) = \mathcal{N}(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_i|^{\frac{1}{2}}} \times \exp\left(-\frac{1}{2}(x-\mu_i)^{Tr} \sum_{i}^{-1}(x-\mu_i)\right) \quad (4)$$

In (4) $\Phi_i = (\mu_i, \Sigma_i)$ represents the parameters for i-th Gaussian density and A^{Tr} is the inversion of matrix A. In general, GMM could be identified with its associated parameters, the parameters are; $\lambda_k = (c_i, \Phi_i, i=1,\ldots,K)$.

4 Database

The emotional speech database which is provided by Berlin University is a standard collection of speech corpus, which is used widely in voice sciences and speech processing scientific resources. This database includes audio recordings of ten actors and actresses (five males and five females) who have pronounced sentences with seven standard classes of emotions in German. These seven classes of emotions include anger, disgust, fear, happiness, neutral, sadness and boredom. In this process, each actor has been asked to express one out of ten predetermined sentences which has more vowels with dedicated emotion (Burkhardt et al., 2005). Approximately 800 recorded sentences are used to prepare this database, and then 500 samples of them selected to choose precisely with respect to emotion recognition by human factors. This method makes it possible to select best sentences which represent the most similar emotions to real natural emotions of speakers with particular emotional states. Also, it performs more accurate recognition with precision higher than 80% and natural selection with more than 60% of choices to increase performance and accuracy of this database (Burkhardt et al., 2005). In this experience we have used 454 enounced emotions with respect to sextet standard emotions which exist in EmoDB.

5 Proposed Approach

The approaches, which commonly are used for speech processing, have derived from the methods that are known as pattern recognition. In particular, each moment of speech signal stream, represents the encoded data which leads to that the analytic works on speech emotion recognition (SER) are closely similar to pattern recognition cycle. To begin with, the words uttered in the input speech signal are analyzed separately and performed the routine to emotion recognition. Then the changes in trend of emotional states determine the prevailed emotional feelings of utterance during the lecture or conversation. This result is performed by the probabilistic filtering method to boost up classification accuracy. The overall view of the proposed approach illustrated in the Figure 1.

5.1 Emotion classification (LGMM)

By using this method, the emotion that is laid in uttered single word, is determined. The main purpose of speech processing by this approach is to recognize emotional states of the speaker, and model the trend of its changes during the long speech. The first level of emotion recognition cycle represents in Figure 2.

At this stage, the pre-processing tasks, including windowing are performed and also silent frames are removed from the input speech signal, and then required features of speech signal are extracted using MFCC method for each single word. In the next step, feature selection is performed to convert obtained coefficients into the required coefficients. This causes to decrease the size of feature vector and prevents curse of dimensionality at the classification process. Then these features are used as an input vector to the classifier. We use a type of Gaussian Mixture Model, which we have modified it to perform learning as the leaning-based GMM. We have called this method as LGMM.

In this paper we propose an expanded derivation of Gaussian Mixture Model to provide classes of emotions using combination of Gaussian densities. The motivation which convinced us to use this type of GMM was that Gaussian components can represent some of spectral shapes of speech signal which depend to general emotions of utterance. Another reason is that the capabilities of Gaussian combinations are so reasonable for stochastic density modeling similar to modeling of speech signals.

To describe mathematically, a Gaussian Mixture Model is generally a weighted sum of several Gaussian components. In other words, Gaussian Mixture Model is a linear combination of M Gaussian densities, which is represented in

$$P(\vec{\mathbf{x}}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{\mathbf{x}}) \quad (5)$$

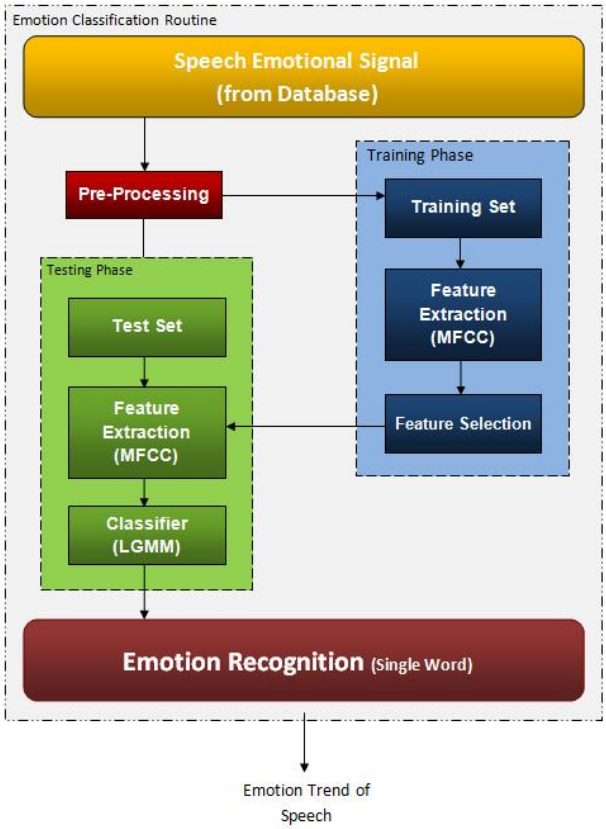


Figure 2. Block diagram of emotion recognition of proposed method for single word of speech.

According to the recent equation, $\vec{\mathbf{x}}$ is a D-dimensional stochastic vector, $b_i(\vec{x})$ are density components for $i=1,...,M$ and p_i are combined weights for $i=1,..,M$. Each component of Gaussian function is D-dimensional and in the form of

$$b_i(\vec{\mathbf{x}}) = \frac{1}{(2\pi)^2|\Sigma^i|^2}\exp\left\{-\frac{1}{2}(\vec{\mathbf{x}}-\vec{\mu}_\mathbf{i})^T\sum_i^{-1}(\vec{\mathbf{x}}-\vec{\mu}_\mathbf{i})\right\} \quad (6)$$

the $\vec{\mu}_\mathbf{i}$ represents the mean vector and Σ_i determines the covariance matrices. Also, combined weights of general probability rule, emphasize the concept that sum of probabilities is equal to 1 and satisfy the main statistical rule which is $\sum_{i=1}^{M} p_i = 1$.

The mathematically flexibility is the prominent advantage of using this method of speech modeling. Intuitively the density of complete Gaussian components can only be shown by mean vectors and covariance matrices. These components are obtained from combination of weights of all density components. Also, probability density functions of destructed features which are affected by differences exist in emotional specifications of those functions. As a result, we could use a set of GMMs to calculate probability of particular emotion which are prevailed by utterance. This method also concludes maximum likelihood estimation which should be determined a class-condition probability density function by providing a Bayesian classifier. For instance, the selection of initial model could be done by using test data, but parameter configuration of this model needs some measures of optimality such as the degree of accuracy when the data distribution is fitted to the observed data. Accordingly, the value of data likelihood is an optimality measure. Just suppose we have a set of independent samples such as $X=\{x_1, x_2,\ldots,x_N\}$ derives from a data distribution which is represented by probability density function like $p(x;\theta)$. In this function the θ is the set of parameters of PDF. The likelihood is represented in

$$L(X;\theta)=\prod_{x=1}^{N} P(x_N;\theta) \quad (7)$$

This equation represents the likelihood of data distribution of X, or in a nutshell, it shows the data distribution of parameter θ. The main purpose of this equation is to find that $\hat{\theta}$ would maximize value of likelihood. We also have in

$$\hat{\theta} = \arg max_\theta = L(X;\theta) \quad (8)$$

This function most often does not reach to its maximum value, but the algorithm mentioned in (9) analytically and mathematically is evident and clear. This equation also called likelihood function:

$$L(X;\theta) = \ln L(X;\theta) = \sum_{n=1}^{N} \ln p(x_N;\theta) \quad (9)$$

Due to uniformity of the logarithm function, a solution that has mentioned in (10) has similar usage to $L(X;\theta)$. According to these definitions, implementation steps of LGMM classifier is as described underneath. At the first point, the parameters are initialized, and then mathematical expectation is taken based on previous probabilities for i=1,…, n and then k=1,…,K are calculated by

$$P_{i,k} = \frac{a_k^{(r)}\emptyset(x_i\mu_k^{(r)},\Sigma_k^{(r)})}{\Sigma_{k=1}^{(r)}a_k^{(r)}\emptyset(x_i\mu_k^{(r)},\Sigma_k^{(r)})} \quad (10)$$

Then maximization likelihood value is provided by

$$a_k^{(r+1)} = \frac{\sum_{i=1}^{n} P_{i,k}}{n} \quad (11)$$

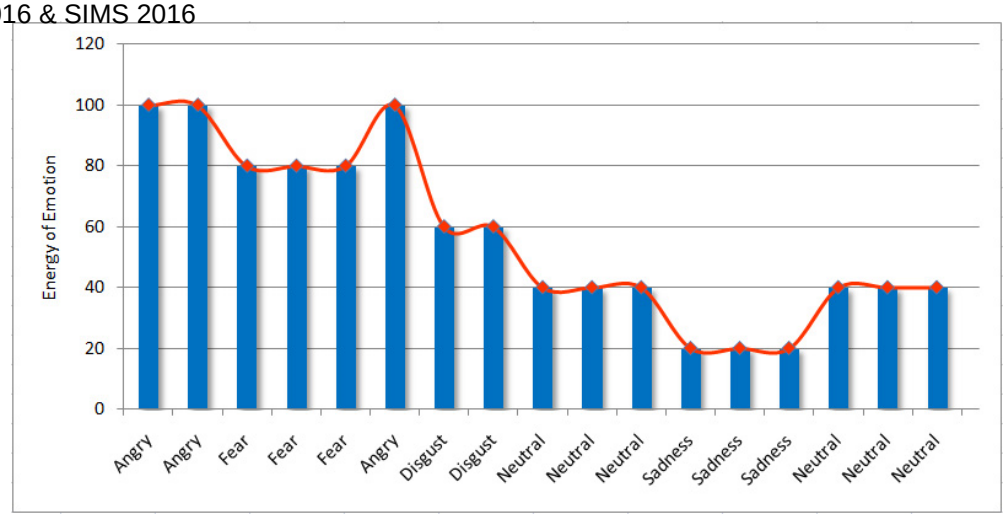


Figure 3. Sample trend diagram of emotion states of utterance during speech.

$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{n} P_{i,k} X}{\sum_{i=1}^{n} P_{i,k}} \quad (12)$$

$$\mu_k^{(r+1)} = \frac{\sum_{k}^{(r+1)} P_{i,k} (x_i \mu_k^{(r+1)})(x_i \mu_k^{(r+1)})t}{\sum_{i=1}^{n} P_{i,k}} \quad (13)$$

And as long as the data converge, steps of getting mathematical expectations and maximizations repeat iteratively. Besides, this data distribution is unknown for us at first, which in the next step; the features are obtained by applying MFCC method. These features are in the form of 12-dimensional space. It is also unknown for us the mode of this data and the number of peaks in its distribution.

So, in this way we begin using a Gaussian component for each emotional class and then calculate parameters. This phase of proposed approach called training phase which the learning tasks take place. Next, each component is divided into two parts and retrained repeatedly for each part, same as classical "divide and conquer" renowned method.

Divisions and trainings continue repeatedly until they reach the final number of required components. Another issue which we had faced using GMM, is that there is not any possible solution to train a Gaussian mixture model with C components (calculation of parameters, $\Sigma, \vec{\mathbf{x}}_\mathbf{i}, p_i$) as a Closed-form equation.

The EM algorithm was used to model the Probability Density Function (PDF) of the emotional speech prosody features in (Schuller, 2004; Lee, 2005). By using this method, optimal Gaussian components are obtained at last in repeated iterations and training task of LGMM performed successfully.

Since we do not have enough data to calculate all parameters of complete covariance matrix, training of GMMs is performed using diagonal covariance matrices. It is also worth noting that the training phase just performs once when the application begins to run.

At this stage of emotional classification, all previously mentioned steps perform on feature vector, which obtained for each single word in uttered speech signal separately. Then the emotion of utterance during expressing the particular word in speech is recognized. At the final stage of this classification level of proposed approach, the labels of emotional classes as the emotional states are obtained to the number of uttered words in the whole speech. These emotional labels could be different for each uttered word due to the changes of emotional states of utterance during the speech or conversation. Finally, we depict the changes trend of the emotional states of utterance during the speech.

5.2 Trend of Changes in Emotional States

Emotional information which is embedded in speech signal is derived from expressed speech input or from parts thereof, this uttered emotion information being descriptive for an emotional state of a speaker and its changes (Kowalczyk and van der Wal, 2013).

Next, we have obtained a trend of emotional changes automatically, during the speech using proposed method. Modeling and simulation of this trend shows the changes of feelings of the speaker when expressing talk or speech. The system could measure changes in emotional states of the utterance by applying the proposed approach. Figure 3 illustrates the trend of an instance speech, which shows the changes of emotional states and moods of the speaker during expressing speech. In Figure 3, it is also

obvious that the mood of the speaker changes between anger, fear, disgust, neutral and sadness during speech.

6 Experiments

In this paper, we propose a method for emotion recognition of utterances during the speech or talk using the extracted features of speech signal. Considered emotional classes are based on standard classification in behavioral and speech sciences. These classes include anger, boredom, disgust, fear, neutral and sadness. Applying the developed method and test it on the data in EmoDB, the results are investigated using Cross-validation method and represented with evaluation parameters in form of accuracy.

To remind, the recognition accuracy is an evaluation method which means how close the measured value to the actual accurate value is. This measure indicates the percentage rate of emotion recognition accuracy for each input speech signal in test phase to total emotional speech data in training phase (Yang and Lugger, 2009). These assessments are provided for each of six emotional states in the domain of emotional classes. The results are obtained based on performance of proposed method on Berlin emotional speech database (EmoDB). The recognition accuracy rates are represented in Table 1.

In consequence, we have calculated analytical and statistical parameters which stem from obtained result. We also do have achieved the score 2.52 as the standard deviation for accuracy rates of the six emotional states. This result shows that our approach represents high degree of stability in emotion recognition from the speech. Also, we have achieved scores, 6.34, 0.0294, 7.42 and 85.50 as variance, dispersion coefficient, variation range and geometric mean, respectively. As well as, the acquired dispersion coefficient also emphasized on the sustainability of system.

Table 1. Recognition accuracy rate on EmoDB.

Emotion	*Classification*
Angry	83.86
Boredom	87.56
Disgust	84.69
Fear	81.32
Neutral	87.08
Sadness	88.74

7 Conclusions

We have demonstrated an approach for speech emotion recognition (SER) and modeling its emotional changes of the speaker using the innovative classification method, which is based on a probabilistic method. To this end, we applied a modified version of GMM as a basis for this approach of emotion classification, which we have named it as Learning Gaussian Mixture Model (LGMM). We have used 12-MFCC to extract features from the raw audio signal of speech. We also have used Berlin emotional speech corpus database (EmoDB) for training and testing the proposed method of emotion recognition. Due to the admissible results in recognition accuracy rates, which are obtained using the proposed method, we do offer this method to design and develop the emotional speech recognition-based systems, specially the systems which need to anticipate human behavior. The main motivation of this research is to simulate the trend of changes in feelings and emotions of speaker during the speech. A prominent advantage using this method is to depict a view of emotional behavior of the speaker regardless to speech context, instant events or factitious behaviors during the speech or conversation. Also, we have applied benefits of using MFCC for feature extraction, which this leads to more accurate results. This method of feature extraction also demonstrates good performance in noisy environments; however, the recognition accuracy could be a little decreased in the very noisy environments. Compared to the conventional methods in the field of emotional speech recognition, despite of the limited number of train and test samples in the database, the obtained results using the proposed method allow us to achieve state-of-the-art consequences in recognition accuracy and run time.

References

X. Anguera, S. Bozonnet, N. Evans, and C. Fredouille. Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing.* DOI: 10.1109/TASL.2011.2125954.

F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss. A database of german emotional speech. *INTERSPEECH*, 12:1517–1520, 2005.

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing magazine*, 18(1):32-80, 2001.

Hesam Farsaie Alaie, Lina Abou-Abbas, and Chakib Tadj. Cry-based infant pathology classification using GMMs. *Speech Communication*, DOI:10.1016/j.specom.2015.12.001, 77:28-52, 2015.

B. Fergani, M. Davy, and A. Houacine. Speaker diarization using one-class support vector machines. *Speech*

Communication, 50:355-365, 2008, DOI:10.1016/j.specom.2007.11.006.

P. Kenny, D. Reynolds, and F. Castaldo. Diarization of telephone conversations using factor analysis. *Selected Topics in Signal Processing, IEEE Journal of*, 4:1059-1070, 2010.

Wojtek Kowalczyk and C. Natalie van der Wal. Detecting Changing Emotions in Natural Speech. *Springer Science and Business Media New York, Appl Intell*, 39:675–691, 2013, DOI: 10.1007/978-3-642-31087-4_51.

Rahul B. Lanjewar, Swarup Mathurkar, and Nilesh Patel. Implementation and Comparison of Speech Emotion Recognition System using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques. *Procedia Computer Science*, 49:50-57, 2015, DOI:10.1016@j.procs.2015.04.226.

C. M. Lee and S. Narayanan. Towards detecting emotion in spoken dialogs. *IEEE transaction Speech and Audio Processings,* 13(2):293-303, 2005.

Sara Motamed and Saeed Setayeshi. Speech Emotion Recognition Based on Learning Automata in Fuzzy Petri-net. *Journal of mathematics and computer science,* 12:173-185, 2014.

B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture. *In Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. (ICASSP '04).*

F. Valente. *Variational Bayesian Methods for Audio Indexing*, PhD. dissertation, Universite de Nice-Sophia Antipolis, 2005. DOI:10.1007/11677482_27.

C.N. van der Wal and W. Kowalczyk. Detecting Changing Emotions in Human Speech by Machine and Humans. *Springer Science and Business Media, NY - Applied Intelligence,* 39(4):675-691, 2013, DOI: 10.1007/s10489-013-0449-1.

D. Ververidis and C. Kotropoulos. Emotional Speech Classification Using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm. *In Proceedings IEEE International Conference on Multimedia and Expo, Amsterdam, 2005.* DOI: 10.1109/ICME.2005.1521717.

L. R. Welch. Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter* 53(1):10-13, 2003.

J. H. Wolfe. Pattern clustering by multi variant analysis. *Multivariable Behavior Res.,* 5:329-359, 1970.

Yang and M. Lugger. Emotion recognition from speech signals using new harmony features. *Special Section on Statistical Signal and Array Processing,* 90(5): 1415–1423, 2010, DOI: 10.1016/j.sigpro.2009.09.009.

3D Virtual Fish Population World for Learning and Training Purposes

Bikram Kawan Saleh Alaliyat

Faculty of Engineering and Natural Science, Norwegian University of Science and Technology, Norway,
bikramkawan@gmail.com, alaliyat.a.saleh@ntnu.no

Abstract

This paper presents the potential use of the 3D virtual world of fish population for training and educational purposes, especially for who are new to fish farming industry. Virtual Reality is the proven technology which is emerging everyday with new methods and implementation. We simulate the fish swimming behavior based on the social rules that are derived from flocking behavior of birds. The simple relation we proposed to represent fish birth and death resembles the biological ecosystem of fish in the sea. The experiment results from different case studies we carried out shows the realistic fish population dynamics. The system user interface gives the users the ability to change the system parameters for different cases to see the real-time effect. Through different case studies carried, our framework can be used to simulate different environments.

Keywords: 3D, virtual reality, fish farming, virtual world

1 Introduction

Technology is emerging and becoming more advanced with the time span. With the development of technology, more and more resources are easy to access from home and school. The distance-learning concept is now proven technology and accepted worldwide. The teaching methodology in school, training and learning methods have been changing in large scale with the use of new technology. One of the most popular technology, which have very great scope in the coming days is to adapt virtual worlds in many sectors including learning and educational sectors. 3D Virtual World (VW), which provides realistic three-dimensional environments and offer engaging, interactive and immersive experiences, creates new opportunities for teaching and learning (Sampson, 2011). These opportunities are related to the realism of the educational activities representation within 3D VWs and to the enhanced aspects of interactivity provided within them. In the recent years, different researchers have recognized the educational and training potential of 3D VWs due to their unique features, such as the recreation of the sense of presence, their immediateness, the real world simulations provided, and the new experiences that may not be possible, non-cost effective and even dangerous to represent in the real world (Molka-Danielsen and Mats, 2009).

In the adult sector large multinationals companies are using virtual worlds to educate their employees, and hospitals (e.g. St George's London), Governments (e.g. Canadian Border Security) and the Military (e.g. USAF MyBase in Second Life and TRADOC in Active Worlds being just two of many examples) are also exploring the use of this technology (Dalen.co.uk, 2010). The need of training simulator is demanding due to cost, safety and environmental purposes. We can use virtual world simulators to model any complex modules in which we can test any danger situation without affecting the life of human and physical damages. Innovative use of virtual reality technologies for the education and training offer new opportunities that can address needs for modular design that are adaptive, safe, flexible and reusable. There have been already another study on the reliability and potential of virtual fish farming as discussed on this paper (Hiemstra, 2015).

This study is primarily for education and training purposes especially in fishing industry. Aalesund is the capital city of fishing industry in Norway. Most of the people are engaged in this business and are willing to do. There are many stages, which needs to be finished in order to step into this industry. People are required to go through several training courses in order to start their fishing business of their own. It is well known that a real life training that involves experiment with alive fish, ocean and environment is quite impractical and not suitable. The testing for training such as what makes the breeding of fish faster, what may cause death of fish, what will happen to the fish environment if we change the parameters of sea (light, quality of ocean water, pollution, ship traffic) will be quite complex and nearly impossible in real life. Our project gives the complete framework to cope with this kind of problems for the fish farmers who are willing to start their own business. The virtual world we have made is completely immersive which plays a great role for giving real feeling to the trainers. In virtual world, they can experiment any parameters and see the effect directly. After going through several hours on training, they will achieve the professional skills which help them to start working in the fish industry.

Our model provides the framework for completely immersive training simulator for fish farmers. The framework uses mathematical modeling such as fish breeding rate, death rate and external factors in terms of mathematical modeling. Steering behaviors of fish such as following leader, changing the heading towards coworkers, maintain the speed of flocks is used to show the movement of fish in a realistic manner (Reynolds, 1999). In this paper, we use the term boids (Boids are bird-like objects that were developed in 1980s to model flocking behavior) to represent the flock of fish. 3D models are used to represent fish, environment, ocean, rocks and other objects to get a virtual world similar to the real life. The materials are used for 3D models, lighting and camera effects to add more realism to the model. In addition, the important part is User Interface (UI) that gives the user easy access to change any parameter such as fish death rate, birth rate, velocity and others that will show direct effect on our simulation. Also, we can add, modify for other complex scenario based on the requirements. Our framework is quite flexible, cost effective and easy to start.

2 Related Work

Virtual Reality (VR) has been studying for many years and is implemented successfully in many different sectors. The first practical use of VR was done by William Winn in 1993 (Youngblut, 1998). Winn discussed the importance of immersive VR technology in three different aspects which are not available in the real world (Winn, 1993). The three aspects he suggested are size, transduction, and reification. In 1994 Jonassen proposed that learners construct their own reality, or at least interpret it based on their perceptions of experiences, so an individual's knowledge is a function of one's prior experiences (Jonassen, 1994). There are plenty of works accomplished on VR in many areas. Some of them are discussed in this section.

Sandra Tan and Russell Waugh mentioned the use of VR in teaching and learning Molecular Biology. She discussed that her project helps students to give a better understanding in visualization for example (DNA) instead of drawing, models and other molecular dynamics on the board. It was easier for students to understand the transcription process in molecular biology (Sandra Tan, 2003).

The company called “minecraftedu” has already implemented Virtual learning in school. They believed this helped kids to explore planets and historical place (Miller, 2012).

Damian Schofield have discussed on effectiveness of advanced three-dimensional (3D) VR technology in chemical engineering and simulates the configuration and operation of a polymerization plant (Scholfield, 2012).

Chih-Kai Huang and his co-authors discussed the implementation of a Virtual Fishing System. They mentioned how the user can incorporate a boat simulator, an interactive fishing rod, and virtual reality fishing scenes. This system creates a spontaneous and interactive environment, and offers the thrill and fun of sea fishing at home or at the amusement park (Chih-Kai *et al,* 2004).

Seungho Park has discussed in his article about the virtual fishing system through digital image sensing. He proposed the system that simulates fishing in a virtual space with changing baits (Park, 2003).

3 Modeling

This section is the core of the whole scheme. It bridges the gap between user interaction and the simulation result directly, which gives the users completely immersive to their result to understand the effect of what they want to simulate.

We have implemented individual agent based system. According to Yndestad (Yndestad, 2015), System is composed of a set of partners collaborating on a common purpose. Agents are related to the landscape. System model in terms of agent-based model can be represented by

$$S(t)=\{A(t), L(t), N(t)\} \quad (1)$$

where, A(t) represents a set of agents, L(t) is a landscape and N(t) is the relationships between the agents them self and between them and the landscape. Yndestad represents the individual agents in his system model as

$$S(\text{Agent},t) = \{A(\text{Arc},t), A(\text{Dyn},t), A(\text{Eti},t), A(\text{Lea},t)\} \quad (2)$$

where A(Arc,t) represents the agent's architecture, A(Dyn,t) is the agent’s dynamics, A(Eti,t) is the agent’s ethics and A(Lea, t) is the agent’s learning.

In our framework, we assumed each fish as an individual agent, ocean as the environment and dynamics such as birth, death, and scaling rates are used to formulate the relations. The agent learning means each fish can adapt its swimming behavior by following steering behavior rules that will be discussed in this chapter in section 3.3. We will describe the model in details in the following sections.

3.1 Fish Modeling

We have divided fish model into fish model and shark model. Also made seven assumptions as in Algorithm I.

3.2 Environmental Modeling

The environment model is very important to give a good realistic system and to simulate the effects of changing the environment parameters on the fish behavior. Some parts of the environment are fixed; others can be changed in the UI directly, while the remainders are changing by simple dynamics equations. We have added trees, bushes and rocks to our model. We have assumed the trees, bushes growing, and decay. In addition, the shifting of rocks is also added. We have defined some rates that are function of time for trees, bushes and rocks objects similar like we defined in the fish model. Besides, we generate a random number and check if this number is less than the rate we define, and then we assume objects are decaying, growing or shifting of rocks.

Algorithm I. Fish and shark modeling.

```
Foreach fish and shark agent
  Initialize {fish birth rate {k1), fish death rate (k2), fish scaling rate
    (k3), shark birth rate (k4), shark death rate (k5), shark scaling rate
    (k6), shark eating fish rate (k7) }
  Calculate probabilities
    {
      Prob. of fish birth (p_k1) = 1/ k1
      Prob. of fish death (p_k2) = 0.00001* k2
      Prob. of fish scaling (p_k3) = 0.00001* k3
      Prob. of shark birth (p_k4) = 1/ k4
      Prob. of shark death (p_k5) = 0.00001* k5
      Prob. of shark scaling (p_k6) = 0.00001* k6
    }
End Foreach

Do
   Generate random number (R)
   Foreach fish agent
     If R < p_k1
           New fish is added
     End If
     If R < p_k2
           Fish is treated as dead and removed from the world
      End If
      If R < p_k3
          The size (length, breadth and height) of fish is      increased
      End If
   End Foreach
   Foreach shark agent
      If R < p_k4
          New shark is added
      End If
      If R < p_k5
           Shark is treated as dead and removed from the world
      End If
      If R < p_k6
          The size (length, breadth and height) of shark is increased
      End If
   End Foreach
   Foreach fish and shark
    If R < k7
          Fish is killed by shark and fish is removed
          New shark is born
    End If
   End Foreach
    While user halt the program.
```

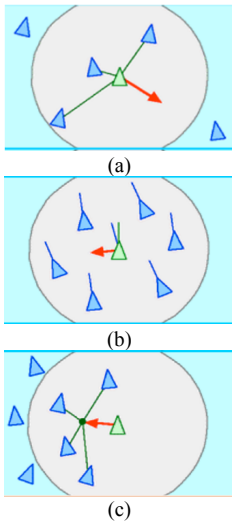
3.3 Steering Behavior Rules

Craig Reynolds proposed the famous steering behavior rules to represent the social behavior movement models in 1996. The basic flocking model consists of three simple social steering behavior rules that describe how an individual boid maneuvers based on the positions and velocities its nearby flock mates (Reynolds, 1999). The three main parameters are summarized below.

Separation: Steer to avoid crowding local flock mates (Figure 1 (a)).

Alignment: Steer towards the average heading of local flock mates (Figure 1 (b)).

Cohesion: Steer to move toward the average position of local flock mates (Figure 1 (c)).



Separation: Steer to avoid crowding local flock mates.

Alignment: Steer towards the average heading of local flock mates.

Cohesion: Steer to move toward the average position of local flock mates

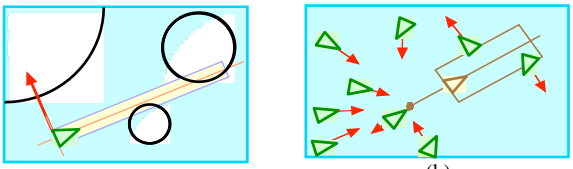
Figure 1. The Reynolds steering behavior rules (Reynolds, 1999).

The proposed steering behavior rules are not enough to adapt the complex environment. Therefore, Reynolds proposed adding individual based rules which will help each individual to adjust steering behaviors even in a complex dynamic environment (Reynolds, 1999). This will eventually help boids to finish specific task. The two important methods he included are:

- Obstacle avoidance (Figure 2(a)): This behavior allows boids to find path against obstacle in cluttered environment situation.
- Leader following (Figure 2(b)): Boids will adapt this behavior to follow the leader for specific task.

We have implemented the steering behavior rules proposed by Reynolds in Unity3D. The program is written in MonoDevelop Unity – C# (www.unity3d.com). The main reason behind implementing this model is to simulate the behavior for

the school of fish. Five rules are implemented in our model (Alaliyat et al, 2014).



(a) (b)

Figure 2. The Reynolds extended steering behavior rules (a) obstacle avoidance (b) leader following (Reynolds, 1999).

Cohesion: This rules is responsible for fish agent to get closer to the center of neighbor fish. Hence the school of fish are formed. This rule can be seen as attraction rule i.e. acts opposite of separation rule. Assume cohesion of boid (b_i) as (Coh_i). It can be calculated in two steps. First we calculate the center of flock (f) represented as ($\overrightarrow{Fc_i}$), which can be obtained by

$$\overrightarrow{Fc_i}=\sum_{\forall bj^{\epsilon f}}\frac{\vec{pj}}{N} \quad (3)$$

where, j is the position of boid j and N is the total number of boids. Then, the gravity of boids to lean toward the center of density of the flock is called as cohesion displacement vector as indicated by

$$\overrightarrow{Coh_i} = \overrightarrow{Fc_i} - \vec{pi} \quad (4)$$

Alignment: This is the rule for boids (fish) to steer towards the same direction with neighboring fishes. There is a certain range of threshold distance of alignment for the fish which are counted as neighbors and average of their velocities are calculated. We can calculate alignment (Ali_i) in two steps. First we find the average velocity vector of neighboring fishes ($\overrightarrow{Fv_i}$) by

$$\overrightarrow{Fv_i}=\sum_{\forall bj^{\epsilon f}}\frac{\vec{vj}}{N} \quad (5)$$

Then displacement vector of Alignment (Ali_i) can be calculated by

$$\overrightarrow{Ali_i}=\overrightarrow{Fv_i} - \vec{v_i} \quad (6)$$

where $\vec{v_i}$ is the velocity vector of boid i without this rule there will be no formation of nice flocking behavior of fish rather than bouncing each other randomly.

Separation: This is the rule for boids to make some distance from neighboring fish. The rule can be seen as type of repulsion rule. It's important to note that the distance from which the boids start to avoid each other must be less than the distance from which the boids attract each other (due to the cohesion rule). Otherwise no flocks would be formed. We have implemented separation (Sep_i)) rule by (7). The self-position of boid b_i and neighboring visible boid b_j vectors are summed together. Then separation steer ($\overrightarrow{Sep_i}$) can be calculated by

$$\overrightarrow{Sep_i} = \sum_{\forall bj^{\epsilon f}}(\vec{p_i} - \vec{p_j}) \quad (7)$$

Leader following: This is the rule to follow nearby moving boid chosen as leader (p_l). The leader following (Led_i)) is calculated by

$$\overrightarrow{Led_i} = L * (\vec{p_i} - \vec{p_j}) \quad (8)$$

where L is a leading strength factor. (Note: the moving vector (velocity) has limits, minimum and maximum.

Random movement: This rule is to add some disturbance in the movement of fish to make realistic. Random number generator of Unity is used for this purpose. The random movement ($Rand_i$)) is calculated by

$$\overrightarrow{Rand_i} = -f\ factor * \vec{r} \quad (9)$$

where r is a unit sphere random vector and f factor is a flock random strength factor.

Then the moving vector (V_i) is for boid (b_i) is calculated by combining all the steering behavior vectors as by

$$\vec{V_i} = w_1 + \overrightarrow{Coh_i} + w_2\overrightarrow{Ali_i} + w_3\overrightarrow{Sep_i} + w_4\overrightarrow{Led_i} + w_5\overrightarrow{Rand_i} \quad (10)$$

where w_i are the coefficients describing influences of each steering rule and used to balance the five rules.

The simulations of fish with all these rules are implemented in Unity3D game engine. The choice of game engine has been more important due to its better visual representation. We imported the 3D model fish, shark and other sea materials from third party instead of wasting time on building 3D models in which we are not interested. The animation such as movement of fish tail and fins make more realistic. All the 3D models are rigid solid object with dynamics of physics. This is important to keep away from collision of objects. We have used box collider and cylinder collider of game engine according to the necessity.

4 3D Modeling and Visualizations

3D model is the presentation of object used in VW which is similar in real life. The basic components of 3D models are cubes, polygons, sphere, cylinder and so on (Cudworth, 2014).

We can model, build, or create whatever we can visualize if that visualization is based on careful, complete observations of what you see in the real world. With scripting, we can change the size, position, color, texture, and visibility of any 3D models and transform them as they are approached and viewed, a great way to keep the experience fresh and interesting.

In the real physical world, we perceived 3D objects and their relationship to each other in space through a variety of "depth cues". Our brains observe the relative size of two objects, and we often assume the larger one is closer to our position in space. By using the time-honored technique of forced perspective by building objects some distance away in a smaller scale, or diminishing these objects in actual scale as they progressively become distant, we can fool the brain into thinking these objects are even farther away (Cudworth, 2014). The 3D models used in our framework are:

4.1 Fish

The 3D model of fish is made of polygon meshes. The fish model has meshes defined with name "fish" and several joints are merged to make one solid 3D fish. It holds the information of animation of fish, i.e. waving the fins when they move. The general procedure of making 3D model fish is similar to shark as shown in Figure 3.

4.2 Shark

The shark model is also a collection of polygons formed in meshes. The mesh information is stored in "humpback" as mesh. The shade used for shark is Bumped Specular. The steps for the formation of shark are shown in Figure 3.

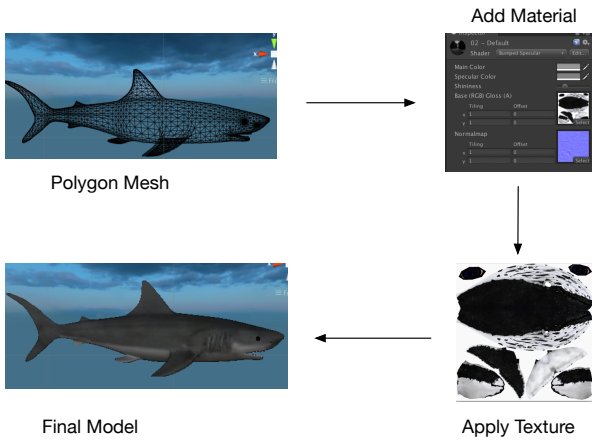


Figure 3. Steps involve in making 3D model of shark.

4.3 Other Models

The other objects models that we have used in our project are:

- Rocks
- Water bubbles
- Water creatures like crab, trees.

4.4 Lighting

There are three basic things that lighting should do within a virtual environment and they are (Cudworth, 2014)

- Illuminate the meaning (or purpose) of this environment
- Support the mood(s)
- Augment the visual style Abbreviations and Acronyms

Even though we have nice textures and structures if there is no light that is meaningless. We have used point light, directional light and spot light.

4.5 Viewing

A Unity scene is created by arranging and moving objects in a three-dimensional space. Since the viewer's screen is two-dimensional, there needs to be a way to capture a view and "flatten" it for display. This is accomplished using Cameras (Unity3d.com, 2016). We have installed several cameras in different areas in the scene, a camera to follow the leader, a camera to visualize the dynamic behavior of trees and rocks and a camera for bushes.

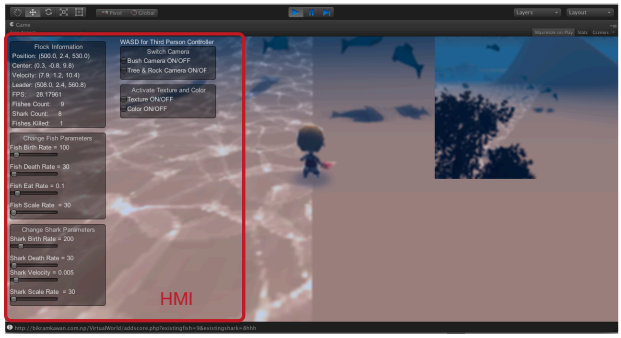


Figure 4. Human Machine Interaction.

4.6 Human Machine Interaction (HMI)

HMI as shown in Figure 4, is necessary to facilitate the user to select various parameters such fish birth rate, death rate, camera selection and so on to see the real time effect on the scene. We have used different types of controlling mechanism for the scene which are as below.

- Label - Display the information related to specific variables.
- Horizontal Slider - Slider has been implemented to facilitate for changing the dynamics of fish and shark populations. It will

easily let the users to change the birth rate, death rate and so on.

- Toggle Button - This is Boolean option for changing the case. It has been implemented for changing the position of camera. For example, the user can easily check the option in the screen to switch the camera from rock area to bushes area or fish flock area.

5 Experimental and Simulation Results

For simulation of fish behavior and other dynamics we have made 3D virtual world in Unity3D game engine. We have made a school of fish based on the steering behaviors rules. The movement of fish can be obtained by (10), and the initial number of fish and shark is shown in Table 1. Figure 5(a), Figure 5(b), Figure 5(c) and Figure 5(d) are the simulation results of steering behavior rules respectively for separation, cohesion, alignment and avoidance.

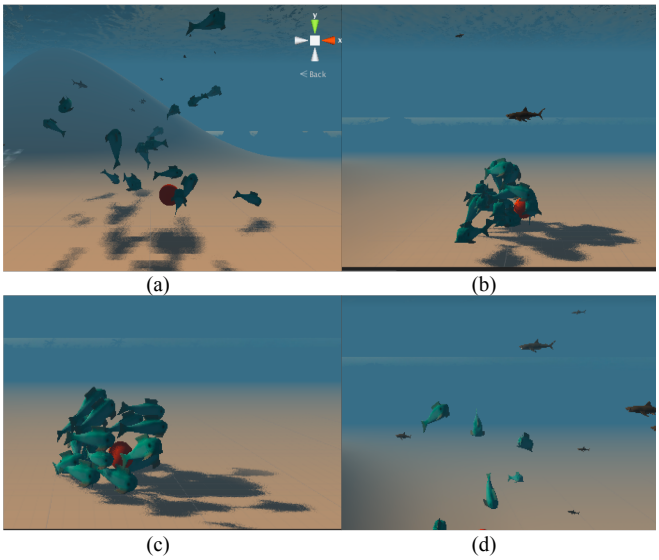


Figure 5. Fish movements' rules. a) Separation b) Cohesion c) Alignment d) Avoidance

Table 1. Parameter Settings.

Name	Minimum	Maximum	Parameter Value
Fish Birth Rate	10	1000	100
Shark Birth Rate	10	1000	100
Fish Death Rate	1	1000	30
Shark Death Rate	1	1000	30
Fish Initial Number	0	N/A	10
Shark Initial Number	0	N/A	5
Eating Rate	0.05	0.5	0.1
Shark Scale Rate	10	1000	30
Fish Scale Rate	10	1000	30
Shark Velocity	0.001	0.05	0.005

An experiment has been carried on the system we have built to verify the feasibility of the proposed flexible structure. The case studies aim to extract and verify the fish behaviors upon the time changing with some parameter settings. The initial parameters' settings before we run the model are shown in following Table 1.

First, we set the parameters as shown in Table 1, and try different changing in the parameters in control setting of the application. We have done some cases as below:

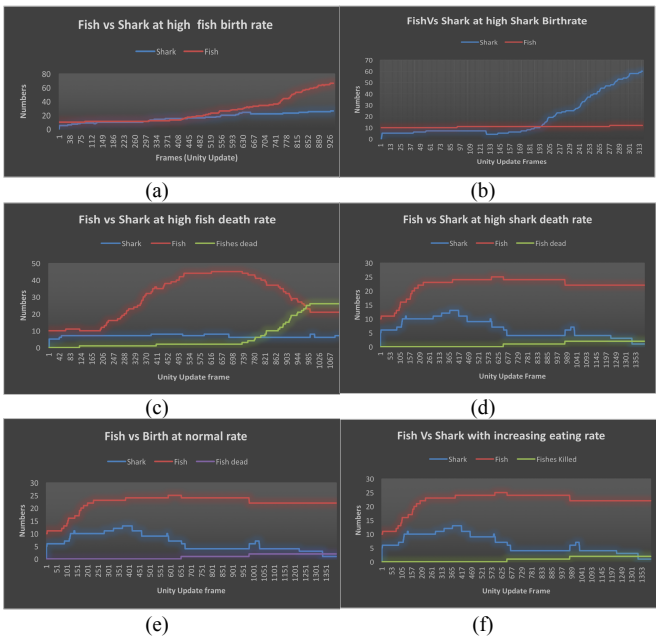


Figure 6. Fish vs Shark on different cases.

Case 1: Changing fish birth rate

We changed the fish birth rate from 100 to 10 making all other parameters constant. The output is shown in Figure 6 (a).

Case 2: Changing shark birth rate

We changed the shark birth rate from 100 to 10 making all other parameters constant. The output is shown in Figure 6 (b).

Case 3: Changing death rate of fish

We increased the population of fish by increasing the fish birth rate and then make fish birth rate very low around 471. The shark birth rate was made low around 921. Later the death rate of fish was shifted from 30 to 926 in slider. We cannot make simulation for very long time due to complexity of scene and rendering freezes. It is shown in Figure 6 (c).

Case 4: Changing death rate of shark

We increase the population of shark by increasing the shark birth rate and then make shark birth rate very

low around 471. The fish birth rate was made constant. Later the death rate of shark was shifted from 30 to 500 in slider. The output is shown in Figure 6 (d).

Case 5: Fish and shark at normal rate

The settings for normal rate, which are set at initial parameter settings, are shown in Figure 6 (e).

Case 6: Fish and shark with eating rate

The eating rate was changed from 0.01 to 0.3 and the results obtained are shown in Figure 6 (f).

6 Conclusions

The inception of VR in education and learning can be tracked back from the 90s. There is lot of potential use of VR in education such as: immersive technology helps student to learn the material, it increases interaction with the world and make feel the user is a part of that world, learners can integrate theoretical ideas by manipulating the 3D object to see how it work. In addition, the learners can create any complex scenario and test them to see the result easily, which is not possible in the real world.

The aim of this paper is to help learners who are interested in fish industry. The learners or students can get a better overview from the system we proposed. The testing of fish behavior in real life is very difficult while it is very simple and easy to understand in 3D virtual worlds by only changing some parameters from the user interface. In addition, learners can get the overview of fish population dynamics based on the setting they have applied in the application with the help of the real time graph. This makes the better understanding when the fish will have high population or less population. Due to the realistic visual representation, users feels completely immersive, which is very important to make the user feel like he is inside a real world.

We have discussed the 3D models we have used in our virtual fish farm, the rules and assumption which we implemented in our system. The simulations of 3D models are extremely helpful for better visual representation and this is facilitating the user correlation with real life. The behavior of fish movement should represent reality which is the main concern of virtual reality as we discussed in the introduction chapter. The overall movement of each individual fish is obtained by summing cohesion, separation, alignment, leader following and random factor as we described in section 3.3 of chapter 3. The fish model such as birth, death, scaling and the probability of getting eaten by shark is another important assumption we implemented in our system in order to understand the ecosystem of fish and shark in the sea. In addition, to make the environment natural-looking; growth function of tree, changing the color of tree to indicate aging and shifting of rock due to sea waves are also implemented in our system. On top of these models, HMI will give easy access to the user to control all the parameters in the system we have developed and to see the effect instantly. Beside these, we have cameras in different section to switch the environment for the user.

The system we proposed here has the necessary modules for education and training of the people who are interested to pursue the fish farming business. The main advantage of using simulation is that the user can test any complex scenario which has zero impact on physical properties. After spending several hours in testing different scenarios it makes better understanding and will increase confident for farmers to start their business.

The challenge in our system is the need of high computing machine to render complex scenario with lots of 3D models. The rendering of the scene become slow if there are many 3D models such as fish and sharks after continues increase in number. Another challenge is the correct tuning in the parameters to resemble the real scenario such as flock formation, death and birth function. If we have a real life mathematical formulation of death, birth of fish and shark it will add more realism to the system. The use of Genetic Algorithm for self-tuning may help to make the system more robust. In addition, many works have worked on integration of devices like Oculus Rift, Google glass and these are compatible with the Unity 3D.

References

S. Alaliyat, H. Yndestad, and F. Sanfilippo. Optimization of Boids Swarm Model Based on Genetic Algorithm and Particle Swarm Optimization Algorithm (Comparative Study). *In Proceeding of European Council for Modeling and Simulation ECMS, Italy,* pages 643–650, 2014, doi:10.7148/2014-0643.

H. Chih-Kai, W. Ming-Shyan, L. Jing, S. Kun-Da, and C. Chia-Ming. Implementation of a virtual fishing system. *In Proceeding of the 2004 IEEE International Conference on Control Applications*, pages 509-514, 2004, doi: 10.1109/CCA.2004.1387262.

A. L. Cudworth. Virtual World Design. *CRC Press*, 2014.

Daden.co.uk. *Virtual Worlds For Education And Training*. Dalen Limited, Birmingham, United Kingdom. [online] available via https://www.daden.co.uk/ [accessed March 20, 2016].

D. G. Sampson. 3D Virtual Worlds in Education and Training. *In Proceeding of IEEE International Conference on Technology for Education, Chennai, Tamil Nadu*, pages 3-3, 2011. doi: 10.1109/T4E.2011.7.

L. Hiemstra. *Virtual Salmon Farming*. Aquaculture North America. N.p. Web. Septemper 08, 2015. Available via https://aquaculturenorthamerica.com/ [accessed March 29, 2016].

D. H. Jonassen. Thinking Technology: Toward a Constructivist Design Model. *Educational Technology*, 34(4): 34-37, 1994.

A. Miller. *Ideas for Using Minecraft in the Classroom*. George Lucas Educational Foundation, 2012. Available via https://www.edutopia.org/blog/minecraft-in-classroom-andrew-miller [accessed March 20, 2016].

J. Molka-Danielsen and D. Mats. *Learning and Teaching in the Virtual World of Second Life*. 1st ed. Trondheim, Tapir Academic Press, 2009, Available via http://urn.kb.se/resolve?urn=urn:nbn:se:oru:diva-56487 [accessed March 20, 2016].

S. Park. Virtual Fishing System through Digital Image Sensing. *Journal of the Asian,* 1: 82-90, 2003.

C. Reynolds. Steering behaviors for autonomous characters. *In Proceeding of Game Developers Conference*, San Jose, *California. Miller Freeman Game Group, San Francisco, California*, pages 763-782, 1999.

D. Schofield. Mass effect: A chemical engineering application of virtual reality simulator technology. *MERLOT Journal of Online Learning and Teaching*, 8(1): 83–78. 2012.

S. Tan and R. Waugh. Use of Virtual-Reality in Teaching and Learning Molecular Biology. *3D Immersive and Interactive Learning*, pages 17-43, 2013. doi.org/10.1007/978-981-4021-90-6_2

Unity3d.com. *Unity - Camera*. [online] Available via http://docs.unity3d.com/Manual/CamerasOverview.html [accessed March 20, 2016].

W. Winn. *A Conceptual Basis for Educational Applications of Virtual Reality.* University of Washington, 1993. Available via https://www.hitl.washington.edu/ [accessed March 29, 2016].

H. Yndestad. *Swarm Intelligence Tutorial*. Norwegian University of Science and Technology, Aalesund, Norway, 2013.

C. Youngblut. *Educational Uses of Virtual Reality Technology.* Institute for Defense Analyses, 1998. Available via https://www.hitl.washington.edu/ [accessed March 29, 2016].

Virtual Reality Simulators in the Process Industry: A Review of Existing Systems and the Way Towards ETS

Jaroslav Cibulka[1] Peyman Mirtaheri[1] Salman Nazir[2] Davide Manca[3] Tiina M. Komulainen[1]*

[1]Department of Electronics Engineering, Oslo and Akershus University College of Applied Sciences (HiOA), Oslo, Norway, tiina.komulainen@hioa.no

[2]Department of Maritime Technology and Innovation, University College of Southeast Norway (USN), Vestfold, Norway, salman.nazir@usn.no

[3]Department of CMIC, Politecnico di Milano, Milano, Italy, davide.manca@polimi.it

Abstract

Simulator training with Virtual Reality Simulators deeply engages the operators and improves the learning outcome. The available commercial 3D and Virtual Reality Simulator products range from generic models for laptops to specialized projection rooms with a great variety of different audiovisual, haptic, and sensory effects. However, current virtual reality simulators do not take into account the physical and psychological strain involved in field operators' work in real process plants. Collaborative training using Extreme Environments Training Simulators could enhance the learning process and provide a more realistic perception of the time and effort needed to carry out demanding operations in Extreme Environments. We suggest developing the following features for an optimal ETS experience and safe learning environment: immersive 3D virtual environments, mixed-reality features, automated assessment, and a monitoring system for the physiological and psychological condition of the trainees.

Keywords: training simulators, extreme environments, condition monitoring, performance assessment

1 Introduction

1.1 Training simulators for process operators

Dynamic process simulators have been used since the 1980s for control room operator training (Cameron, 2002; Nazir and Manca, 2014). In Norway, the government requires simulators to be used in operator training (Petroleum Safety Authority Norway, 2012). The benefits of operator training simulators are well documented (Ayral and Jonge, 2013; Cheltout et al., 2007; Fiske, 2007; Komulainen et al., 2012; Nazir et al., 2015; Sneesby, 2008). Typically, simulator training is organized with one instructor for a team of 2-5 control room operators. The simulation scenarios include for example shut-downs, start-ups, abnormal situations, and new procedures (Komulainen and Sannerud, 2014). In these scenarios, the simulator instructor plays the role of field operator and performs all the necessary actions instead of the field operator using the instructor computer. Operator training with high-fidelity control room simulators represents the most effective transfer of knowledge and skills: there is a strong correlation/overlap with reality and the simulation situation is almost identical to reality. Thus, simulator training effectively increases trainees' learning outcome. (Kluge et al., 2014; Spetalen and Sannerud, 2013; Tuomi-Gröhn and Engeström, 2003)

One of the aims of Virtual Reality (VR) is to evoke photo-realistic and immersive feelings and emotions that ensure the user is deeply engaged and involved in the simulation scenario. There is a close relationship between physical perception and learning capability (Bergouignan et al., 2014), (Medical Xpress, 2014). The physical manifestation of feelings enables us to store new information and improve our ability to memorize and remember. It has been proven that emotions are linked to memory processes and thus improve both learning and training (Neuroscience News, 2015), (Tendler and Wagner, 2015). The more intense the emotion, the stronger the record saved in the memory. Furthermore, stress emotions stimulate memory and the learning retention rate. Hence, VR simulators are considered an effective training tool. It has been experimentally demonstrated that VR training is more effective than conventional lectures with figures and videos (Nazir et al., 2013). Trainees memorize hands-on practice, even virtual, better than learning in classrooms by means of slides or videos (Nazir et al., 2015). Field operators can use VR simulators to train routine operations, procedures, abnormal conditions, start-ups, shut-downs, and emergency situations.

A combination of Operator Training Simulators (OTS) and Virtual Reality Simulators is used to train communication skills and team work between control

room operators and field operators (Colombo et al., 2014; Manca et al., 2013; Nazir et al., 2014). Control room operators use a dynamic process simulator (OTS) that is connected to the Virtual Reality Simulator used by the field operators. The Virtual Reality Simulator is an immersive environment where the work site is represented in 3D with stereoscopic vision and spatial sounds (Manca et al., 2013).

1.2 Extreme environment training simulators (ETS)

Current VR simulators do not take into account the physical strain and effort involved in field operators' work in real processing plants. The environment and wearing protective gear increase the field operator's physical workload, and to some extent impair his /her perception (Manca et al., 2016; Nazir et al., 2015). Collaborative training using an Extreme Environment Training Simulator (ETS) can enhance the learning process and can give a more realistic perception of the time and effort required to carry out demanding operations in extreme environments.

1.3 Research questions

This paper surveys and discusses recent use of 3D and VR simulators for operator training in the process industry and team training between control room operators (CROP) and field operators (FOP). We propose the following research questions:

What kind of Virtual Reality Simulators are available to the chemical process industry? Which features are included in the Virtual Reality Simulators? Which features should be developed in order to provide physically and psychologically realistic operator training in Extreme Environments?

1.4 Research methodology and scope

The literature search included companies that provide 3D/VR simulators for the chemical process industry. Other industries were outside of the scope of the search. The literature study is mainly based on the product datasheets available on the companies' websites.

2 Review of current 3D and virtual reality simulators

In the following, we review the main features of nine different 3D/Virtual Reality Simulators that are available to the chemical process industry. A summary of this review is presented in Table .

The companies and their 3D/VR products included in this study are:

- EON Reality: I3TE - Immersive 3D Operator Training Simulator (EON Reality, 2015; World Oil, 2015)
- GSE Systems: Activ3Di (GSE Systems, 2013; GSE Systems, 2014)
- Illogic: VR Star (Illogic, 2016)
- Kairos 3D: Gilgamesh (Kairos3D, 2016)
- MMI Engineering: QUARTS - Quantitative Real Time Hazard Simulator (MMI Engineering)
- Schneider Electric: SimSci-EYESIM (Schneider Electric, 2015; Schneider Electric Software, 2015)
- Siemens: COMOS Walkinside ITS (Siemens AG, 2013; Siemens AG, 2016)
- Simtronics: Virtual Field Operator (VFO) (Simtronics, 2016)
- Simulation Solutions: 3D Virtual Reality Outside Operator (Simulation Solutions Inc., 2016)

2.1 VR projection tool and audiovisual immersiveness

The visual features of the virtual reality plant are projected to the user by means of a computer screen, a pair of 3D glasses, a head-mounted-display (HMD) or a VR projection room. The virtual reality audio can be provided by computer speakers, a stereoscopic headset or the audio system in a VR projection room. An example of the projection room is shown in Figure 1.

Computer screens with loudspeakers are inexpensive and computationally less demanding than other alternatives. Five of the companies included in this study use 3D glasses or a head-mounted-display such as the Oculus Rift (Oculus VR, 2015).

The VR projection room is a quite expensive solution, although there are many different alternatives on how to implement such rooms, also known as Cave, Dome or Cube (Muhanna, 2015). Only one of the companies included in this study uses a VR projection room –EON Reality with its EON Icube(EON Reality, 2016).

2.2 VR avatar control

The movements of the trainee in the virtual reality environment, e.g. the avatar, can be controlled using a keyboard, a mouse, a gamepad, voice commands, or a gesture tracking device (Leap Motion, 2016).

Gesture-tracking devices used in Virtual Reality solutions include Microsoft Kinect, Leap Motion's Orion (Leap Motion, 2016) and VICON Bonita B10.

2.3 Features: Immersiveness

An immersive system evokes photo-realistic and immersed feelings/emotions of being deeply engaged and involved. A good immersive effect can be achieved by using affordable, high-quality head-mounted-displays combined with directional and surround sound effects. Immersive Virtual Reality training is more effective than conventional training.

Simulators that use computer screens for visualization, loudspeakers for sound, and a keyboard or mouse for avatar control are not Immersive Virtual Reality simulators, but 3D simulators.

Simulators that use more advanced hardware tools can be classified as VR simulators, the degree of immersivity depends on the different visual/audio/haptic effects projected to the user (EON Reality, 2015).

2.4 Features: View and multi-user capabilities

All of the 3D and VR simulators only have a first person (1p) view, but two of the products have a third person view (3p).

Most of the Virtual Reality simulators can have multiple users in the same scenario, which is very realistic for many scenarios including emergency situation training.

2.5 Features: 3D objects and graphic effects

All of the VR products allow the trainee to interact with the equipment and to see the status / position of the equipment in the VR plant, such as valve opening and measurement values. Four of the VR producers have implemented seamless motion effect for the equipment changing from one state to another. This feature increases the transfer outcome between the simulated situation and the workplace situation, and improves the learning effect.

Many of the VR simulators allow the user to see through the equipment in the VR environment, for example the liquid level in the tanks. Two of the VR simulators provide visualization of fire and smoke, and use CFD-like simulations to predict the spread of gases and liquids, see Figure 2.

2.6 Features: Augmented Virtual Reality (AVR)

Another feature that some of the advanced VR simulators have adopted recently is Augmented Virtual Reality. As the name suggests, additional information is imposed in the virtual environment to improve the user's learning. Figure 3 represents this feature.

2.7 Features:immersive 4D sensory effects

The EON reality I3TE reports enhanced 4D immersive sensory effects, including tactile feedback, odors, vibration, and wind simulation (EON Reality, 2015). The vibration is generated by a motor under the floor, and the wind is generated by directional fans.

2.8 Features: Learning management system

Some of the 3D/ VR simulators include a learning management system (LMS) that gives an automatic score to the trainees after the simulation scenario. The learning outcome can thus be quantitatively assessed. A VR simulator with LMS can be used for (periodic) verification of the operators' competencies. This feature enables the user/trainee to continuously improve his/her performance. Generally, the assessment of performance is made by an expert assessor, which can easily be biased by his/her experience and knowledge. Therefore, objective performance measures ensure more robust, consistent, and unbiased assessments of performance.

Figure 1: Process simulator by TSC Simulation is connected to the virtual reality environment I3TE by EON reality.

Figure 2: Emergency situation, Visualization of fire in COMOS Walkinside simulator (Siemens AG, 2013).

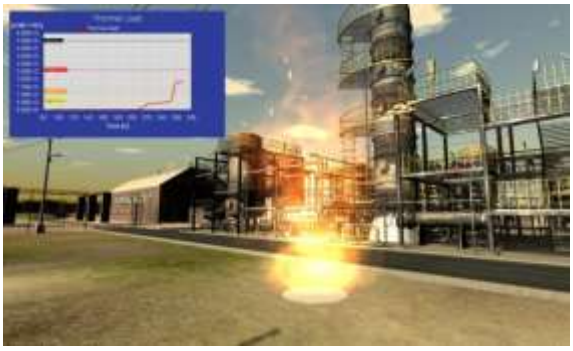


Figure 3: A simulation of an accident where AVR is used to impose additional information (graph on left top corner) (Nazir and Manca, 2014).

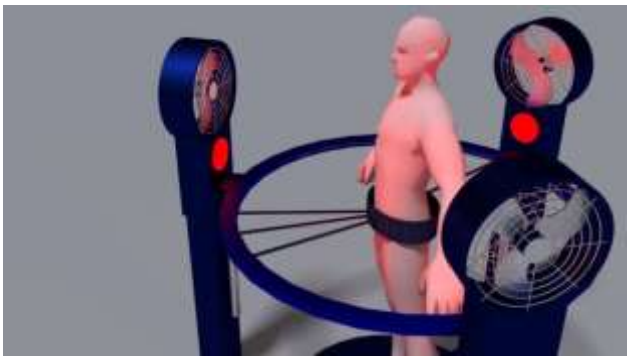


Figure 4: Omnidirectional treadmill with embedded IR lamps and fans.

3 Extreme environment training simulator (ETS)

The aim of the ETS is to improve the immersiveness of VR simulators by including all of the physical strain and effort involved in field operators' work in real processing plants. The environment and wearing protective gear increase the physical workload of the field operator, and to some extent impair his/her perception (Manca et al., 2016; Nazir et al., 2015). In the following subchapters, we summarize the current VR features required by the ETS simulator, and propose features that need to be developed together with the required safety and health measures

3.1 Existing features

The existing features in VR simulators include:

• Audio effects: noise, equipment-specific sound patterns, process-specific sound patterns, weather-specific sounds such as wind

• Visual effects: light, weather patterns, spread of gases and liquids. Fire, smoke, assembly/disassembly of equipment, see- through (X-sec) view

•Visual effects in Augmented Reality: Thermal/Radiation/Toxic exposure

• Sensory/ olfactory effects: odors (low H2S).

• Sensory/ haptic effects: wind (directional fans), vibration feedback (motor under floor)

• Sensory/ thermal effects: heat

• Collision detection (optical motion capture)

• Learning management system

3.2 Features to be developed

In addition to the existing VR simulator features, the following features should be developed in order to provide fully immersive training sessions (Manca et al., 2016). A greater degree of immersion is expected to improve Key Performance Indicators (KPIs) and Distributed Situation Awareness (DSA).

- Environmental factors that need to be implemented in the ETS in a safe manner: temperature, pressure, gravity, elevation/height.
- Environmental factors requiring protective gear: harsh conditions, low oxygen, high exposure to radiation, acids/bases/salts, water deficiency, polluting/toxic substances.

We propose the following equipment be used in ETS sessions:

- Audiovisual equipment: Smartphone VR headsets represent an attractive and low-cost alternative to the self-contained PC HMDs. Utilizing smartphones is a versatile, mobile and untethered VR solution. For VR rendering, it employs the smartphone's display, GPU, CPU, memory, sensors, and other embedded features. The embedded sensors, i.e. MEMS gyroscope and accelerometers, are used for head tracking and data logging.
- Physical equipment representing the distance and elevation changes in the processing plants: the Steward platform or the Omnidirectional treadmill (Virtuix, 2016) presented in Figure 4. Enables the trainee to walk, run, crouch, jump, and freely rotate. It has embedded motion detection sensors, including foot tracking in the low-friction baseplate. The trainee is safely secured in the harness as part of the revolving ring.
- Weight compensation representing gravity and elevation: Adjustable weight compensation such as the Counterbalance weight system in the Cyberith Virtualizer (Virtual Reality Reporter, 2015).
- Wearables representing the work tools, protective gear including suit, shoes, gloves glasses, masks, and oxygen bottles.
- Haptic gloves providing feedback from the interactive process equipment and work tools in the VR environment, for example CyberTouch II (CyberGlove, 2016).
- Heat could be simulated to the trainees: Infrared lamps could be used as presented in Figure 4 (Cyberith, 2016) or overalls/garments with embedded HVAC features.
- Wind could be simulated to the trainees: Directional fans as presented in Figure 4 (Cyberith, 2016).
- Pressure could be simulated to the trainees: Overalls/garments with embedded HVAC features.
- Augmented reality could be used to visually compensate for the health related factors such as low oxygen, water deficiency, polluting/toxic substances. These could be implemented as visual effects like blurry vision, black spots, flickering.
- The learning outcome should be assessed using automatic assessment system with EE performance indicators as described in (Nazir et al., 2015).

3.3 Safety and health measures for EE training

The trainee is safely secured in the revolving harness that surrounds the treadmill (Virtuix, 2016) presented in

Figure 4. The cables from the VR headset and other equipment are secured by an overhead boom.

Due to the hard physical strain of the Extreme Training Simulation, the health status of the trainees should be monitored. We propose monitoring vital signs such as:

- Body core temperature (headphones)
- Heart rate (pulse watch)
- Respiration rate

In addition to parameters such as:

- Eye movement tracker (headset)
- Balance (accelerometer and magnetometer)
- EMG (strain on the legs)
- Pulseoximeter (oxygen saturation)
- Stress indicators such as Cortisol or algorithms to calculate the stress level

Normal values for the core temperature vary between 36.7 and 37.5 (Sund-Levander et al., 2002). Several non-invasive thermopile sensors can measure the temperature of various parts of the body and predict the core (intragastric) body temperature, and thus indicate the heat stress in industrial applications (Graveling et al., 2009). These sensors can be integrated into the headphones in order to measure the temperature within the ear canal, or be placed in the headbands or integrated into the clothing. The heart rate and respiration can be measured by current chest band devices. The maximum heart rate is usually calculated as 220 minus age. There are several studies that have reported heart rate variability in correlation with stress, see for example (Thayer et al., 2012). Regarding the stress level, the respiration rate may increase rapidly during stress, which causes hyperventilation. During hyperventilation,, oxygen saturation increases while the level of arterial CO_2 decreases. In general, a decrease of CO_2 in the blood will consequently decrease the diameter of the blood vessels, including the main blood supply to the brain. This reduction in the blood supply to the brain leads to symptoms such as lightheadedness and tingling in the fingers. Severe hyperventilation can lead to loss of consciousness.

Monitoring these parameters continuously during training will provide feedback on the physiological, and, indirectly, psychological state of the trainees, which is important for their safety but also for optimizing the learning effect.

4 Conclusions

The available commercial 3D and Virtual Reality Simulator products range from generic models for laptops to specialized projection rooms with a great variety of different audiovisual, haptic and sensory effects. Over the past few years, VR simulators have capitalized on the technology advancement in an optimal manner, such that newly integrated features have improved their effectiveness/usefulness.

However, Extreme Environment effects such as temperature, pressure, gravity, elevation/height, harsh conditions, low oxygen, high exposure to radiation, acids/bases/salts, water shortage, polluting/toxic substances should be developed in order to provide realistic training sessions. Monitoring physiological and psychological conditions based on the trainees' vital signs and stress-related factors should be included in order to ensure a safe training environment. In combination with the performance assessment indicators, these parameters may also be applied as markers for an optimal simulator experience.

This article paves the way for research on the evolving topic of Extreme Environment Training Simulators.

Acknowledgements

The authors acknowledge the European Union's Erasmus+ program for funding the teacher mobility that led to the research collaboration on Extreme Environment Training Simulators. The Norwegian Labour and Welfare Administration is gratefully acknowledged for funding the first author.

References

T. Ayral and P. D. Jonge. Operator training simulators for brownfield process units offer many benefits. *Hydrocarbon Processing*, 92(2):45-47, 2013.

L. Bergouignan, L. Nyberg, and H. H. Ehrsson. Out-of-body–induced hippocampal amnesia. *Proceedings of the National Academy of Sciences*, 111(12):4421-4426, 201 4. doi: http://dx.doi.org/10.1073/pnas.1318801111

D. Cameron, C. Clausen, and W. Morton. Dynamic simulators for operator training. In B. Braunschweig, Gani, R., editor, *Software Architectures and Tools for Computer Aided Process Engineering*. Volume 11, pages 393-431, 2002

Z. Cheltout, R. Coupier, and M. Valleur. Capture the long-term benefits of operator training simulators. *Hydrocarbon Processing*, 86(4):111-116, 2007.

S. Colombo, S. Nazir, and D. Manca. Immersive virtual reality for decision making in process industry: experiment results. *SPE Economics & Management*, 2014.

CyberGlove. *CyberTouch II*. Available via http://www.cyberglovesystems.com/cybertouch2/ [accessed 2.6.2016, 2016].

Cyberith. *Cyberith Virtualizer*. Available via http://cyberith.com/ [accessed 1.6.2016, 2016].

EON Reality. *Immersive 3D Training Environment (I3TE) - Enhanced Virtual Reality Based Training for Plant Operators/Engineers*. Available via http://www.eonreality.com/?wpfb_dl=54 [accessed 5.2.2016, 2015].

EON Reality. *I3TE - Immersive 3D Training Environment*. Available via

http://www.eonreality.com/applications/i3te/ [accessed 5.2.2016, 2016].

T. Fiske. Benefits of dynamic simulation for operator training. *Hydrocarbon Processing*, 86(12):17-17, 2007.

R. Graveling, L. MacCalman, H. Cowie, J. Crawford, and P. George. *Reliable industrial measurement of body temperature: The use of infrared thermometry of tympanic temperature to determine core body temperature in industrial conditions.* Institute of Occupational Medicine, 2009.

GSE Systems. *GSE Systems Introduces 3Di-TouchWall, Affordable Alternative to CAVE Simulation for Immersive 3D Training*. Available via http://www.gses.com/news/gse-systems-introduces-3di-touchwall-affordable-alternative-to-cave-simulation-for-immersive-3d-training [accessed 31.05.2016, 2013].

GSE Systems. *GSE Activ3Di Trailer*. Available via https://vimeo.com/89492818 [accessed 31.05.2016, 2014].

Illogic. *VR Star virtual reality platform*. Available via http://www.illogic.us/#!/vr-star-virtual-reality-platform [accessed 17.2.2016, 2016].

Kairos3D. *Kairos3D interactive 3D solutions*. Available via http://www.kairos3d.it/solutions/gilgamesh/ [accessed 31.05.2016, 2016].

A. Kluge, S. Nazir, and D. Manca. Advanced Applications in Process Control and Training Needs of Field and Control Room Operators. *IIE Transactions on Occupational Ergonomics and Human Factors*, 2(3-4):121-136, 2014. doi: 10.1080/21577323.2014.920437

T. M. Komulainen and R. Sannerud. Survey on simulator training in Norwegian oil & gas industry, in *HIOA rapport* vol. 4, ed. Oslo: Oslo and Akershus University College,, p. 27, 2014.

T. M. Komulainen, R. Sannerud, H. Nordhus, and B. Nordsteien. Economic benefits of training simulators. (in English), *World oil*, (12):R61-R65 2012.

Leap Motion. *Leap Motion Introduces Orion, Its Next-Generation Hand Tracking Product for Developers in VR/AR*. Available via https://www.leapmotion.com/news/leap-motion-introduces-orion-its-next-generation-hand-tracking-product-for-developers-in-vr-ar [accessed 29.2.2016, 2016].

D. Manca, S. Brambilla, and S. Colombo. Bridging between Virtual Reality and accident simulation for training of process-industry operators. *Advances in Engineering Software*, 55:1-9, 2013. doi: http://dx.doi.org/10.1016/j.advengsoft.2012.09.002

D. Manca, S. Nazir, T. M. Komulainen, and K. I. Øvergård. How Extreme Environments Can Impact the Training of Industrial Operators. *Chemical Engineering Transactions*, 52:6, 2016.

Medical Xpress. (2014, 10.3.2014) Outside the body our memories fail us. *Medical Xpress*. Available: http://medicalxpress.com/news/2014-03-body-memories.html

MMI Engineering. *MMI Engineering - QUARTS Hazard Simulation*. Available via http://www.atticusdigital.com/mmi-engineering-quarts-hazard-simulation [accessed 16.2.2016,

M. A. Muhanna. Virtual reality and the CAVE: Taxonomy, interaction challenges and research directions. *Journal of King Saud University - Computer and Information Sciences*, 27(3):344-361, 2015. doi: http://dx.doi.org/10.1016/j.jksuci.2014.03.023

S. Nazir and D. Manca. How a plant simulator can improve industrial safety. *Process Safety Progress*, 2014. doi: 10.1002/prs.11714

S. Nazir, S. Colombo, and D. Manca. Testing and analyzing different training methods for industrial operators: An experimental approach. *Computer Aided Chemical Engineering*, 32:667-672, 2013.

S. Nazir, A. Kluge, and D. Manca. Automation in Process Industry: Cure or Curse? How can Training Improve Operator's Performance. *Computer Aided Chemical Engineering*, 33:889-894, 2014. doi: http://dx.doi.org/10.1016/B978-0-444-63456-6.50149-6

S. Nazir, L. J. Sorensen, K. I. Øvergård, and D. Manca. Impact of training methods on Distributed Situation Awareness of industrial operators. *Safety Science*, 73:136-145, 2015. doi: http://dx.doi.org/10.1016/j.ssci.2014.11.015

S. Nazir, D. Manca, T. M. Komulainen, and K. I. Øvergård. Training Simulator for Extreme Environments. *In Proceedings of the* Creating Sustainable Work-environments NES2015, Lillehammer, Norwegian Society of Ergonomics and Human Factors 2015.

Neuroscience News. (2015, 7.8.2015) Emotions Directly Influence Learning and Memory Processes. *Neuroscience News*, . Available: http://neurosciencenews.com/learning-memory-emotion-limbic-system-2393/

Oculus VR. *Start Building: Start turning your dreams into virtual realities.* Available via https://developer.oculus.com/ [accessed 19.1.2016, 2015].

(2012). *Veiledning til aktivitetsforskriften Til § 23 Trening og øvelser*. Available: http://www.ptil.no/aktivitetsforskriften/category383.html#_Toc345589379

Schneider Electric, SimSci - Solutions for Enhancing Refinery Performance and Profitability, ed: Schneider Electric,, 2015.

Schneider Electric Software. *EYESIM Immersive Training System*. Available via http://software.schneider-electric.com/pdf/datasheet/eyesim-immersive-training-system/ [accessed 5.2.2016, 2015].

Siemens AG. Powerful 3D visualization with COMOS Walkinside , 2013.

Siemens AG. *Training for Safety and Profit - Virtual 3D Immersive Training Simulator Makes Case for Stronger ROI*. Available via https://webservices.siemens.com/adtree/newsdb/detail/html.aspx?language=en&filename=TOTAL%20EP

_COMOS_2014_en.xml&frame=1&view=0&design=1&print=1 [accessed 31.05.2016, 2016].

Simtronics. *Virtual Field Operator (VFO) - Interactive 3D Virtual Training Environment*. Available via http://www.simtronics.com/site/vfo.htm#.V01frK1f1aQ [accessed 31.05.2016, 2016].

Simulation Solutions Inc. *Simulator Software - Generic Training Simulators*. Available via http://simulation-solutions.com/simulator-software.html [accessed, 2016].

M. Sneesby. Operator training simulator: myths and misgivings. *Hydrocarbon Processing*, 87(10):125-127, 2008.

H. Spetalen and R. Sannerud. Erfaringer med bruk av simulering som transferstrategi. (in Norwegian), *Nordic Journal of Vocational Education and Training*, 3:17, 2013.

M. Sund-Levander, C. Forsberg, and L. K. Wahren. Normal oral, rectal, tympanic and axillary body temperature in adult men and women: a systematic literature review. *Scandinavian Journal of Caring Sciences*, 16(2):122-128, 2002. doi: 10.1046/j.1471-6712.2002.00069.x

A. Tendler and S. Wagner. Different types of theta rhythmicity are induced by social and fearful stimuli in a network associated with social memory. *eLife*, 4, 2015. doi: http://dx.doi.org/10.7554/eLife.03614

J. F. Thayer, F. Åhs, M. Fredrikson, J. J. Sollers Iii, and T. D. Wager. A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, 36(2):747-756, 2012. doi: http://dx.doi.org/10.1016/j.neubiorev.2011.11.009

T. Tuomi-Gröhn and Y. Engeström. In *Between school and work: new perspectives on transfer and boundary-crossing*. Amsterdam: Pergamon, pp. X, 333 s., 2003.

Virtual Reality Reporter. (2015, 1.7.2015) Cyberith Virtualizer: Immersive VR Gaming Equipment. *Virtual Reality Reporter*. Available: https://virtualrealityreporter.com/cyberith-virtualizer/

Virtuix. *Virtuix Omni*. Available via http://www.virtuix.com/ [accessed 1.6.2016, 2016].

World Oil. (2015, 5.5.2015) ExxonMobil awards license to EON Reality for Immersive 3D Operator Training Simulator technology. *World Oil*. Available: http://www.worldoil.com/news/2015/5/05/exxonmobil-awards-license-to-eon-reality-for-immersive-3d-operator-training-simulator-technology

Table 1. Comparison of VR simulator products.

VR company	EON Reality	GSE Systems	Illogic	Kairos 3D	MMI Engineering	Schneider Electric	Siemens	Simtronics	Simulation Solutions, Inc.
VR Product	I3TE - Immersive 3D Operator Training Simulator	Activ3Di	VR Star	Gilgamesh	Quantitative Real Time Hazard Simulator - QUARTS	SimSci-EYESIM	COMOS Walkinside ITS	Virtual Field Operator (VFO)	3D Virtual Reality Outside Operator
Learning Management System	✓	✓	N/A	✓	N/A	✓	✓	N/A	✓
Immersive VR	✓	×	✓	×	✓	✓	✓	×	×
VR projection	room (EON Icube), 3d glasses, HMD (Oculus Rift, Samsung Gear VR)	×	monitor, 3d glasses, HMD (Oculus Rift)	×	monitor, HMD (Oculus Rift)	3d glasses/stereoscopic headset, 3D projection, 3D HDTV	monitor, 3d glasses, HMD (Oculus Rift)	×	×
Avatar control	gesturing (VICON Bonita B10), voice commands, gamepad	keyboard+mouse, gamepad	gamepad	gamepad	keyboard+mouse, gamepad	gloves, gesturing, gamepad	gesturing (via Kinect), keyboard+mouse, gamepad	keyboard+mouse, gamepad	mouse
Viewpoint	1p	1p	1p	1p	1p	1p, 3p	1p, 3p	1p	1p
Multi-user training	✓	✓	✓	✓	N/A	✓	✓	N/A	✓
Interactive 3D objects	valves, buttons, gauges, etc.	✓	valves, gauges, tanks, engines, etc.	✓	✓	✓	live-action items	N/A	valves, pumps, controllers
4D Immersive effects	3D sound, tactile feedback, odors (H_2S), vibration (under floor), 3D wind (fans)	N/A	N/A	visual effects, item-specific sounds/noises	N/A	N/A	3D sound, odors, vibration	N/A	×
Graphic effects	N/A	see through/in equipment, X-sec view	N/A	see through/in equipment, X-sec view	Fire/Smoke - CFD, adjustable model	see through/in equipment, X-sec view	Fire/Smoke - CFD	N/A	✘
Enhanced features	N/A	N/A	weather conditions	assembly/ disassembly of equipment	thermal radiation exposure, fall risk, injury/fatality level	PPE selection, compatibility with mobile devices,	radiation/toxic exposure, RFID chips (FOP location)	N/A	✘

Recognizing Steel Plate Side Edge Shape automatically using Classification and Regression Models

Pekka Siirtola[1] Satu Tamminen[1] Eija Ferreira[1] Henna Tiensuu[1] Elina Prokkola[2] Juha Röning[1]

[1]Biomimetics and Intelligent Systems Group, P.O. BOX 4500, FI-90014, University of Oulu, Oulu, Finland
{pekka.siirtola, satu.tamminen, eija.ferreira, henna.tiensuu, juha.roning}@ee.oulu.fi
[2]SSAB Europe, Raahe plate mill, Finland

Abstract

In the steel plate production process it is important to minimize the wastage piece produced when cutting a mother steel plate to the size ordered by a customer. In this study, we build classification and regression models to recognize the steel plate side edge shape, if it is curved or not and the amount of curvature. This is done based on time series data collected at the manufacturing line. In addition, this information needs to be presented in a way that enables fast analysis and long-term statistical monitoring. It can then be used to tune the parameters of the manufacturing process so that optimal curvature can be found and the size of the wastage piece can be reduced. The results show that using the classification and linear regression methods, the side edge shape can be recognized reliably and the amount of curvature can be estimated with high accuracy as well.

Keywords: steel manufacturing, classification, regression, plate plan pattern control, plate side edge

1 Motivation

In the steel plate production process, the molten steel is first converted into slabs which still are hot and glowing. Then the thickness of the slab is reduced in the reversible multi-pass rolling process. Plate rolling can be divided in three main stages: sizing, broadside rolling and finishing rolling. In sizing stage the slab is rolled in longitudinal direction to produce required intermediate thickness. Before broadside rolling the slab is turned around 90 degrees and then rolled in transverse direction to obtain the required plate width. After broadside rolling the slab is turned again 90 degrees and rolled to final thickness (Ginzburg, 1989). After rolling the customer plates are cut from the mother plate. As the steel material and manufacturing are expensive, it is desired that the amount of cutting wastage is as small as possible. The uneven shapes at the plate end sides and lateral sides cause yield loss, amounting to about 5% to 6% of a total tonnage of slab used (Ruan et al., 2013). To minimize this loss, the shape of the rolled mother plates needs to be optimized.

During the rolling process, inhomogeneous plastic deformation occurs as width spread at the plate edge regions. There happens width spread at the plate side edge portions, which forms convex shape at the plate ends. Laterally material spreads more at the both plate ends compared to central portion thus having a tendency towards concave plate side edge shape. However, because there is also broadside rolling phase in addition to the longitudinal rolling phase, the plate shape is dependent on the combination of the longitudinal rolling ratio and the broadside rolling ratio. The bigger the broadside rolling ratio is compared to longitudinal rolling ratio, the more convex the final plate side edge shape is formed. Width spread at the plate edge regions is the prime cause of the uneven shapes formed during the hot rolling process (Ruan et al., 2013). There are techniques designed to produce a true rectangular shape of rolled plate. One of the methods is MAS-rolling which was developed by Kawasaki Steel Corporation (Yanazawa et al., 1980).

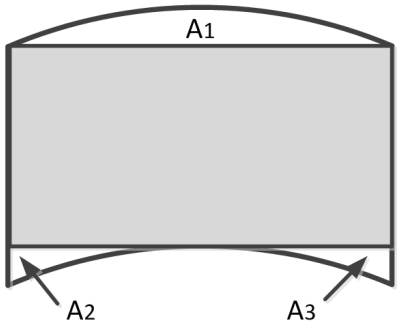


Figure 1. Wastage of concave side is much smaller than of convex side with same curvature but opposite direction, $A_1 > A_2 + A_3$.

However, due to rolling inaccuracies the rectangular shape cannot always be achieved even with MAS-rolling. There is width deviation not only between plates but also inside a plate. This results sometimes in convex and sometimes in concave plate side edge shape. However, the amount of wastage from convex side is much bigger than from concave side with same curvature, see Figure 1. In addition, concave shape increases the size of a rectangular shaped plate in camber shaped plates. Usually the width accuracy is as its best in the middle of the plate and weakens towards the ends. For this reason the deviations in the

ends are bigger and thus more extra material need to de designed there, which also favors concave shaped plate. Therefore, in order to minimize the amount of wastage, a slightly concave edge shape of a steel plate should be preferred. Using MAS-rolling, the target plate side edge shape can be designed slightly concave in order to almost completely avoid plates with convex edge shape. The desired amount of curvature can be defined by adjusting the process parameters. However, because of the uncertainty related to the steel plate production process, exactly the desired amount of curvature cannot be reached. In order to optimize the MAS-rolling parameters and thus curvature of a plate, the shape of the mother plates needs to be monitored.

Our study is made for SSAB Europe, Raahe plate mill, Finland, where the monitoring of the plate shape is currently done visually at the cooling banks by own eyes. There is no camera based monitoring system available and if plates that have already passed cooling banks needs to be viewed, plate shape can be visualized by means of collected process data. Plate shape information can be gathered from thickness gauge data, see Section 3 for more details. However this data is difficult to analyze making visual monitoring time consuming. In addition, conclusions made using visual monitoring are always based on subjective view. Furthermore, unlike our approach, this type of monitoring does not allow the modelling of statistical distribution of the curvature based on historical production data, which can be used to understand the uncertainty of the manufacturing process and to optimize process parameters to obtain the optimal curvature of plates. If the plate shape can be optimized, the mother plates can be manufactured by using smaller slabs which leads to better yield.

In this study, we build models to define the shape of a time series describing the steel plate side edge, if it is curved or not and the amount of curvature, in such form that they can be analyzed in a glance and enable long-term statistical monitoring. The article is organized as follows: related work is covered in Section 2 and the used data set is described in Section 3. Our method is introduced in Section 4, and the method is validated in Section 5. We will discuss our results in Section 6 and, finally, the conclusions are in Section 7.

2 Related work

Many aspects of steel manufacturing have been modelled and optimized using machine learning and data mining methods which have been used for instance to estimate impact toughness (Tamminen et al., 2010), to diagnose faults (Tian et al., 2015), to model the yield strength (Koskimäki et al., 2007) and to model the rolling temperature (Tiensuu et al., 2011).

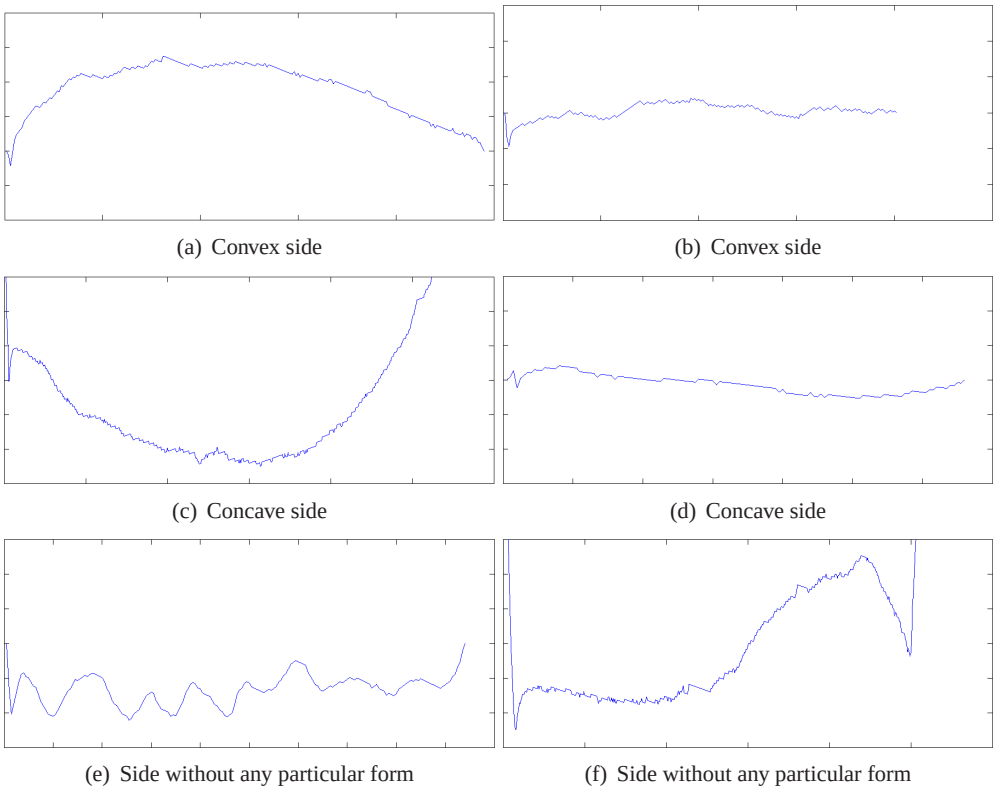
In this study, we aim to define the curvature of a time series describing the steel plate side edge, and use this information to build statistical distribution model to visualize what kind of curve shapes the studied data set includes and how the amount of curvature is distributed in the manufacturing process. In the literature the term *plate plan pattern control* is used to describe techniques and methods that are designed to ensure that produced steel plates have the desired shape and to minimize wastage (NIIR Board of Consultants & Engineers, 2006). Most of the studies related to plate plan pattern control concentrate on improving the rolling process by installing new equipment to the manufacturing line making them expensive, such as Zhang et al. (2015); Inoue et al. (1988), and there are not many studies where plate view pattern is improved by tuning the process parameters based on the data collected in the manufacturing line as we do in our study.

Lee et. al. introduced a neural network based method to predict the width of the mother steel plate based on the size of a slab and different manufacturing parameters (Lee et al.). Their method can be used to select parameters to rolling process so that finished steel plate has the desired width. Juutilainen *et. al.* predicted the steel strip's width rejection probability with statistical models and optimized the working allowance of the product based on the results, material and rejection costs (Juutilainen et al., 2012, 2015). While the width prediction of a steel plate is closely related to our study, our approach differs from it in a way that we study steel plates one side at a time. Therefore, our study gives better understanding about the plate plan pattern than studies which concentrate on the width of a plate and, therefore, study both sides simultaneously.

The studies by Ruan *et. al.* (Ruan et al., 2014, 2015) are the closest to our study. In Ruan et al. (2014) regression analysis was used to build models to predict plate plan view pattern based on slab size and rolling parameters. Models were trained based on simulation results and validated using industry tests. The study concentrates on analyzing what causes convex and concave sides. In our study we are not trying to predict the shape of a side or find causes for certain shapes from the data collected during the rolling process. Instead, our aim is to get better understanding about the manufacturing process by monitoring the ratio of different shapes and the amount of curvature from the plain view patterns measured from finished steel plates. This enables better understanding about the uncertainty related to the manufacturing process and how process parameters should be selected to minimize the yield loss caused by uncertainty.

3 Data set and pre-processing

The data were collected from steel plate production line at SSAB Europe, Raahe plate mill, Finland. The measurements were done using a thickness measurement system from IMS Systems Inc. This measurement system contains three radiation sources and radiation detectors. When a steel plate moves through the production line, the amount of detected radiation can be used to measure not only the thickness of the plate but also its width. In this



(a) Convex side (b) Convex side

(c) Concave side (d) Concave side

(e) Side without any particular form (f) Side without any particular form

Figure 2. The data were classified into three classes: convex, concave, and other. This figure shows examples of members of different classes.

study these width measurements are considered as a time series that describes the shape of a steel plate side edge. The distance between two adjacent measurements is approximately 50 mm, so for an average plate of 20 m, each side can be described as a time series containing 400 measurements.

The data were collected from 399 plates, and therefore, from 798 sides. The length of the plates varied from 4.0 meters to 27.5 meters and width from 1.3 meters to 3.3 meters. Based on this data, the aim of this study was to build models to detect the direction of curvature and the amount of curvature of a curve describing the side of a steel plate. However, the data set did not include labels or correct values about the amount of curvature. Therefore, in order to build a model to recognize the shape of a steel plate side edge, it was necessary to label the data set by hand. When the data was visualized to label it, it was noted that there are many cases where it is difficult to say whether the curve describing the side is convex or concave because plates can be over 20 meters long but still width within the plate only differs a couple of millimeters. In addition, sides are not regular or symmetric, and they can contain features from both convex and concave shapes. It was also noted that all edges are not convex or concave, instead, due to uncertainty and errors in the manufacturing process sides can have other shapes as well. Eventually, the data were labeled into three classes: 'convex', 'concave', and 'other'. There were 160 convex sides, 556 concave sides and 82 instances classified as 'Other'. Examples of instances from each class are shown in Figure 2. Note that because the data is hand-labeled it can contain errors as in some cases it was difficult to conclude what is the correct class label for the side.

Before the data were analyzed, pre-processing was done. In the pre-processing stage, the measurements describing the head and tail of the side edge were removed by removing 10 percent of the measurements from both ends. This was done as typically the head and tail of the side are rounded, and therefore, they are not indicative of the shape of the rest of the side. In addition, when the width of the steel plate is measured, the plate is not necessary positioned fully straight at the manufacturing line. This skewness caused by biased positioning was removed from the measurements in the pre-processing stage by straightening the time series describing the side of a steel plate based on the line described by the first and last measurement point of the side.

4 Defining the curvature of a steel plate

Recognition of the direction and the amount of curvature of a curve describing the steel plate side edge is basically a regression problem as the goal can be considered as one continuous value where the sign of the value tells the direction of curvature and absolute value describes the

amount of curvature. However, it is very difficult to visually estimate the amount of curvature and give an accurate continuous value of describing it. Therefore, the amount of curvature of time series T describing a side edge of a steel plate was estimated using the following equation:

$$curvature_{estim}(T) = \frac{abs(max(T) - min(T))}{\|T\|}. \quad (1)$$

This means that curvature estimation was done by calculating the amplitude of a time series T describing the steel plate side edge, and dividing this with the length of a steel plate. When this information was combined with class labels; defining whether the plate is convex, or concave; we build a response vector where the sign of a response tells the direction of the curvature, and the absolute value tells the amount of curvature. By applying regression analysis to this response vector and features extracted from the measurements it was possible to built a linear regression model to be used to define the shape of a plate, and to define the curvature of a plate.

In addition to convex and concave sides, there are also sides that are neither of these two options. Therefore, these need to be recognized from the data set before regression model can be trained. In this study, this was done using a binary classifier that classifies instances into two classes: the side is either curve (convex or concave) or not.

As a conclusion, the recognition of the direction and the amount of curvature of the side of a steel plate is divided into two tasks: (1) Classify instances into two classes: side is either an curve or not, and (2) Using data from convex and concave side's, train a regression model using estimation of the amount and direction of curvature as response.

In both tasks, the models were trained using the same feature set consisting of the following features: polynomials of degree 1, 2, and 4 were fitted to the time series describing the side and the obtained coefficients as well as the error of the fitting were used as features. Fitting was done not only to the whole time series but also different sizes of parts of it, for instance to the first and the second half separately. In addition, a straight line was fitted through the first and the last measurement of the time series and the ratio describing the number of points below and above this line was used as a feature. This was also done to different parts of the time series so that altogether 23 features of this type were extracted. Other features included minimum, maximum, and different percentiles of the values. The complete feature vector included 66 features.

5 Experiment

In this section, the method presented above is applied to the data set introduced in Section 3. As it was stated in the previous section, the recognition of the direction and the amount of curvature of a steel plate side edge is divided into two tasks: at first it was recognized whether the side was convex/concave or not. In the second phase, convex and concave sides were further studied to estimate the amount of curvature as well as the direction of the curvature.

Table 1. Recognizing non-curve side edges.

Classifier	Accuracy	Precision	Recall
QDA	94.6%	87.0%	82.5%
LDA	91.5%	77.1%	78.9%
C4.5	93.4%	86.2%	73.4%
kNN, k=1	93.9%	87.1%	76.0%
kNN, k=3	94.9%	91.5%	77.9%
kNN, k=5	95.0%	93.2%	78.8%

Table 2. Confusion matrix showing the classification results using QDA.

	Curve	Other
Curve	699	17
Other	26	56

In the first part, the classification was done using four different methods to compare their performance. The classifiers used were *k*NN, LDA, QDA and C4.5. The idea of the *k* nearest neighbor classifier is to classify a data point into the class, to which most of its *k* nearest neighbors belong. In this study, *k* values 1, 3 and 5 were employed. Linear discriminant analysis (LDA) is used to find a linear combination of features that separate the classes best. The resulting combination may be employed as a linear classifier. QDA (quadratic discriminant analysis) is a similar method, but it uses quadratic surfaces to separate classes. C4.5 is a decision tree model that based on the difference in entropy partitions the space spanned by the input variables to maximize the score of class purity. This is done so that the majority of points in each cell of the partition belong to one cell (Hand et al., 2001).

In order to avoid over-fitting in the classification process, the data set were randomly divided into five parts and cross-validation was performed so that one part at the time was used for testing and the rest for training. The most descriptive features from the feature set were selected using a sequential forward selection (SFS) method.

The results are shown in Table 1. In the table, *accuracy* is defined as *true positives / all*, *precision* was calculated as *true positives / (true positives + false positives)* calculated for both classes and averaged , and *recall* stands for *true positives / (true positives + false negatives)* calculated for both classes and averaged. In addition, the results using QDA are shown in more detail in Table 2.

In the second part, a linear regression model was trained using the same feature set to estimate the direction and amount of curvature of the plate's side. Curvature values calculated using Equation 1 combined with information about the direction of the curvature were used as a response vector to this regression model. Table 3 demonstrates the accuracy of recognizing the direction of curve

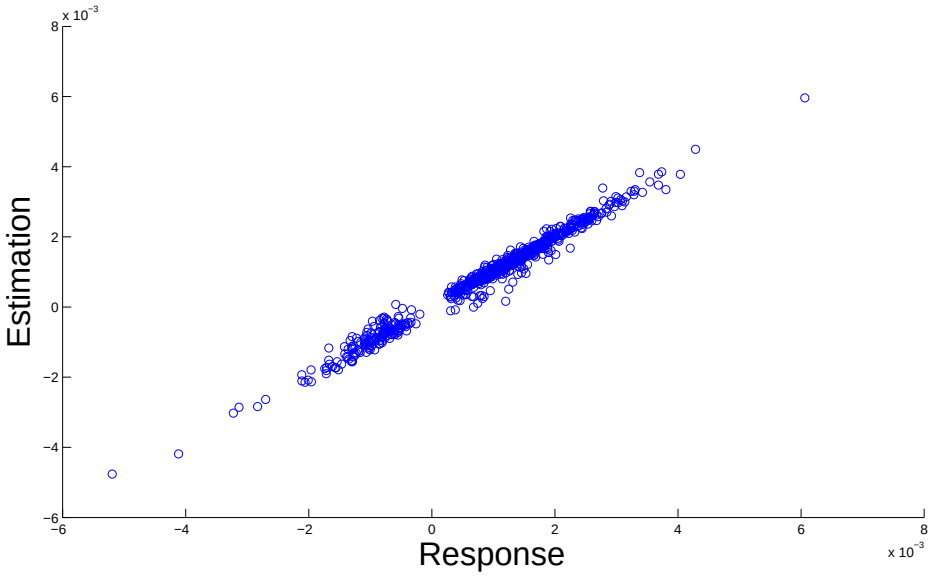


Figure 3. Using linear regression model the direction and curvature of a side can be estimated accurately.

Table 3. Three methods to detect the direction of the curvature of a side are compared.

Method	Accuracy
Regression	99.6%
Intuitive	96.5%
Classification, QDA	99.6%

curvature using the trained model. The results of regression model were compared to two alternative methods: *intuitive method* and *classification* using QDA. Intuitive method is simple, a line is drawn between the first and last observation of the time series, and the direction of curvature is decided based on which side of the line has more observations. In the case of QDA, 5-fold cross-validation was applied and side edges were classified as curve or not.

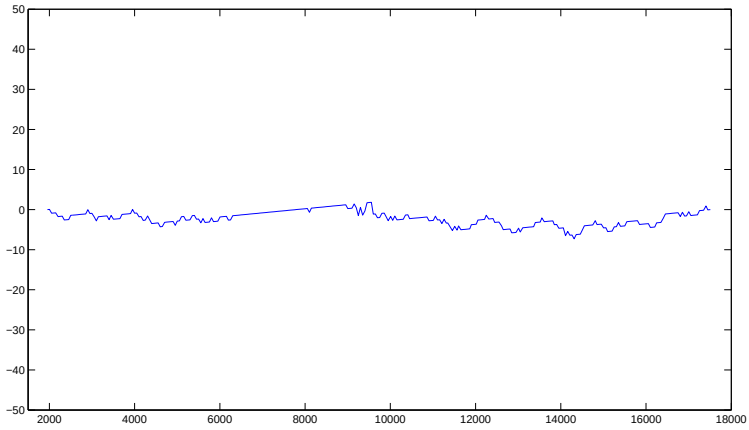
Finally, Figure 3 shows how well the estimation given by the regression model is comparable to the response calculated using Equation 1.

6 Discussion

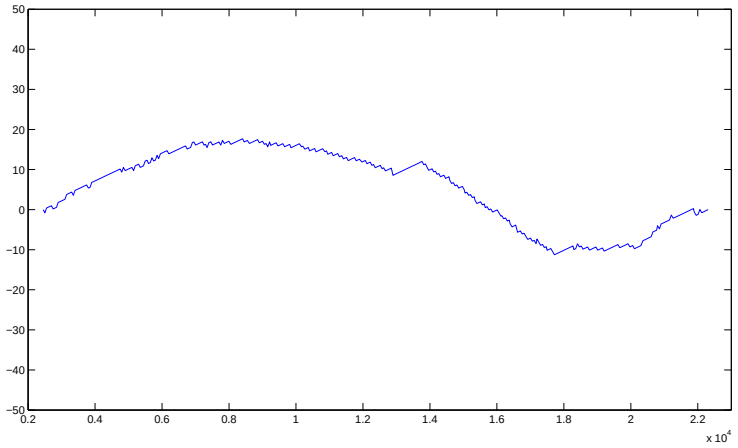
In our approach, the recognition of the shape of the side of the mother steel plate was divided into two phases: at first we recognized whether the side was convex/concave or not. In the second phase, convex and concave sides were further studied to estimate the amount of curvature as well as the direction of the curvature. The recognition rates in Table 1 show that sides without a curve shape can be detected with high accuracy. All the tested classifiers performed well, only LDA had a bit lower recognition rates, although it could still recognize shapes with more that 90 percent accuracy. While the classification accuracy was very high, the precision and recall were not. This is because class 'Other' had a lot lower recognition accuracy than class 'Curve'. This is partly due to the nature of the data set. It can be seen from Table 2 that the data set is very unbalanced. There are a lot more convex/concave sides than ones without any particular form. From manufacturing perspective this is of course good but from a model training perspective it is problematic because there is not the same amount of data from both classes.

Table 3 shows how well the direction of curvature can be detected. It can be seen that using linear regression model it can be detected almost perfectly, in fact in 714 cases out of 716 the curve direction was recognized correctly. Therefore, the method is very reliable. Detection of the direction of the curvature using the QDA classifier performs equally well, while the intuitive method gives a somewhat lower recognition accuracy. However, the advantage of the regression model compared to the classification is that it does not only detect the direction of curvature but it also estimates the amount of curvature, see Figure 3. However, the accuracy of the estimation of the amount of the curvature is more difficult to validate as the response used to train the regression model was also an estimation. In addition, the estimation calculated using Equation 1 is very vulnerable to anomalies of the data which can be caused for instance by errors in the measurements. In this sense, the estimation given by the regression model can be considered more stable and reliable as it is not based on only the measurements data of one side of a steel plate but on the whole data set including 798 sides. Therefore, anomalies of single sides do not have that much effect on the estimation given by the regression model.

Figure 4 shows two sides where the estimation given by the regression model differed the most compare to estimation obtained using Equation 1. In Figure 4(a) the amount



(a) Curvature of this side based on regression model is 1.1mm, and based on Equation 1, 9.1mm.



(b) Curvature of this side based on regression model is 17.4mm, and based on Equation 1, 26.7mm.

Figure 4. Sides where curvature estimations differ the most.

of curvature based on regression model is 1.1mm, meaning that thinnest point of the steel plate side edge is 1.1 mm narrower that the widest point. On the other hand, according to Equation 1 this difference is 9.1mm. In this case, Equation 1 has over-estimated the curvature because in this case there is a lot of variation in measurements making estimation based on amplitude inaccurate. Therefore, the estimation given by regression model, which suggests that side is almost straight, can be considered more reliable in this case. Also in Figure 4(b) the estimation given by regression model can be considered more reliable that the estimation obtained using Equation 1 which has over-estimated the curvature because of the drop at the end of the end of the time series.

Our results show that using the presented method the shape of the side edge of a mother steel plate can be detected with high accuracy and the results can be presented in a way that enables analyzing a steel plate in a glance and enable long-term statistical monitoring. For example, our method to detect sides without curve shape can be used to monitor the manufacturing process and detect if there are possible problems in the manufacturing process. Figure 5 shows the classified data set and sides without curve shape are drawn with higher spikes. Now based on this visualization it can be seen that there are a intervals containing a lot of sides without curve shape, one for instance around x-axis value 40. These types of intervals can be a consequence of some unexpected change in the process, and monitoring enables a quick response to it. In addition, using the estimations obtained using the regression model, in Figure 6 the deviation of convex and concave sides is visualized. Negative value means that side edge is convex, bigger negative value stands for more convex side edge. Similarly, positive value means that side edge is concave and zero equals to straight side edge. Figure 6 shows that majority of the sides are concave but there is still quite many convex plates. Based on these statistics it is possible to define new parameter values for the rolling process to minimize the number of convex plates and to minimize the cutting wastage.

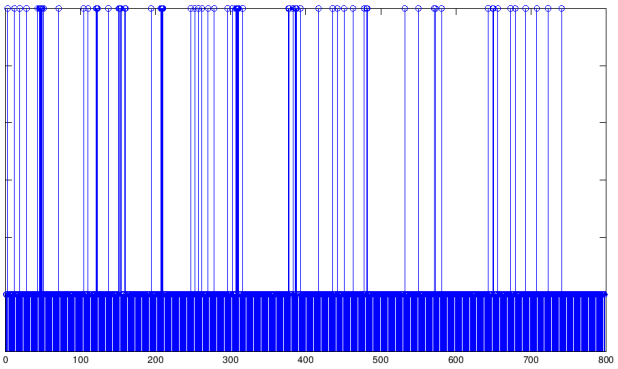


Figure 5. Many of the sides without curve form are centered at clusters. Short bars describe sides with curve form and long without.

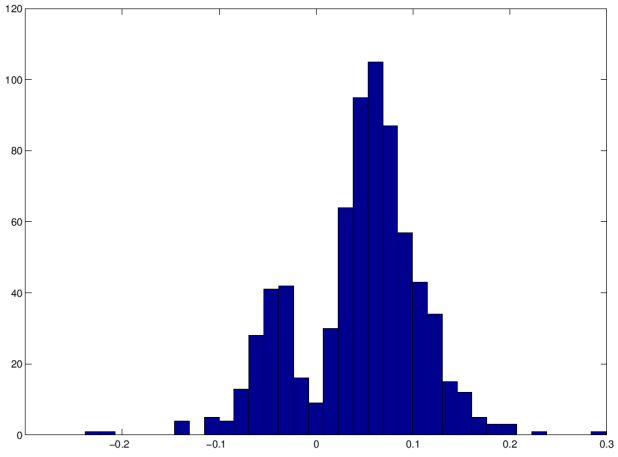


Figure 6. Curvature estimation can be used to optimize rolling parameters. In this case, rolling parameters should be tuned to reduce the number of convex sides.

7 Conclusions

In this study we developed a method to define the steel plate side edge shape based on data collected at the manufacturing line. The method consists of two phases: at first we recognize whether the side is a curve or not, and them we estimate the amount of curve curvature as well as the direction of the curvature. According to our experiments, classification methods can accurately be used in the first phase to detect if the side is a curve. In addition, our experiments show that linear regression model can be used to accurately recognize not only the direction of the curve curvature but also to estimate the amount of curvature.

Acknowledgment

The authors would like to thank SSAB Europe, Raahe plate mill, Finland for providing the data and their expertise for the study. Further acknowledgments are given to the Finnish Funding Agency for Technology and Innovation (TEKES) and Infotech Oulu for supporting this research.

References

V. Ginzburg. *Steel-rolling technology: theory and practice*. Marcel Dekker, Inc., 1989.

D. J. Hand, H. Mannila, and P. Smyth. *Principles of data mining*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0-262-08290-X.

M. Inoue, K. Ohmori, T. Orita, I. Okamura, S. Isoyama, and M. Tarui. Development of a process for manufacturing trimming free plates. *Transactions of the Iron and Steel Institute of Japan*, 28(6):448–455, 1988.

I. Juutilainen, S. Tamminen, and J. Röning. A tutorial to developing statistical models for predicting disqualification probability. In *Computational Methdos for Optimizing Manufacturing Technology, Models and Techniques*, pages 368–399. IGI Global, 2012.

I. Juutilainen, S. Tamminen, and J. Röning. Density forecast based failing probability predictors in manufacturing. *European Journal of Industrial Engineering*, 9(4):432–449, 2015.

H. Koskimäki, I. Juutilainen, P. Laurinen, and J. Röning. Detection of the need for a model update in steel manufacturing. In *Proc. of the Fourth Internation Conference on Informatics in Control, Automation and Robotics (ICINCO 2007), Angers, France*, pages 55–59, 2007.

D. Y. Lee, H. S. Cho, and D. Y. Cho. A neural network model to determine the plate width set-up value in a hot plate mill. *Journal of Intelligent Manufacturing*, 11(6):547–557.

NIIR Board of Consultants & Engineers. *Steel Rolling Technology Handbook*. ASIA PACIFIC BUSINESS PRESS Inc., 2006.

J. H. Ruan, L. W. Zhang, S. D. Gu, J. Zhang, W. He, and S. H. Chen. Establishment of models for plan view pattern control in heavy plate rolling process based on 3-d fem simulation. *International Journal of Materials and Product Technology*, 47(1-4):103–125, 2013.

J. H. Ruan, L. W. Zhang, S. D. Gu, W. B. He, and S. H. Chen. Regression models for predicting plate plan view pattern during wide and heavy plate rolling. *Ironmaking & Steelmaking*, 41(9):656–664, 2014.

J. H. Ruan, L. W. Zhang, Z. G. Wang, T. Wang, Y. R. Li, and Z. Q. Hao. Finite element simulation based plate edging model for plan view pattern control during wide and heavy plate rolling. *Ironmaking & Steelmaking*, 42(8):585–593, 2015.

S. Tamminen, I. Juutilainen, and J. Röning. Quantile regression model for impact toughness estimation. In *Advances in Data Mining. Applications and Theoretical Aspects*, pages 263–276. Springer Berlin Heidelberg, 2010.

Y. Tian, M. Fu, and F. Wu. Steel plates fault diagnosis on the basis of support vector machines. *Neurocomputing*, 151, Part 1:296 – 303, 2015. ISSN 0925-2312.

H. Tiensuu, I. Juutilainen, and J. Röning. Modeling the temperature of hot rolled steel plate with semisupervised learning methods. In *Proc. 14th International Conference on Discovery Science, Lecture Notes in Computer Science*, volume 6926, pages 351–364. Springer, October 5-7 2011.

T. Yanazawa, J. Miyoshi, K. Tsubota, T. Ikeya, H. Kikugawa, and K. Baba. Development of a new plan view pattern control system in plate rolling. *Kawasaki steel technical report*, pages 33–46, 1980.

T. Zhang, B. Wang, Z. Wang, and G. Wang. Side-surface shape optimization of heavy plate by large temperature gradient rolling. *ISIJ International*, advpub, 2015.

Comparison of Different Models for Residuary Resistance Prediction

Elizabeta Lazarevska

Faculty of Electrical Engineering and Information Technologies - Skopje, University Ss. Cyril and Methodius - Skopje, Macedonia, elizabeta.lazarevska@feit.ukim.edu.mk

Abstract

The paper presents several unconventional models of residuary resistance based on fuzzy logic and neural network techniques. First, two fuzzy models are built based on different hull parameters and different Froude numbers. These models are identified by a modification of Sugeno and Yasukawa identification algorithm. Next, a neuro-fuzzy model of residuary resistance is build, based on statistical learning theory. The model presents a fuzzy inference system of Takagi and Sugeno type that uses an extended relevance vector machine for learning its parameters and number of fuzzy rules. Finally, a neural network approach is applied to build four different models of residuary resistance. Two of the neural models apply classic extreme learning machine, and the other two implement incremental extreme learning machine philosophies. The obtained models are validated for their generalization and approximation performance, and although they all possess excellent approximation capabilities, our neural models based on extreme learning machine have shown the best simulation results.

Keywords: residuary resistance, fuzzy modeling, neuro-fuzzy model, extreme learning machine, random nodes

1 Introduction

Obtaining a hydrodynamic model of a sailing yacht is an important step in its initial design, because the model can be used for calculation of the most important hydrodynamic forces acting upon the yacht, evaluation of yacht performance and estimation of its required propulsive power. Within these efforts, Delft Ship Hydromechanics Laboratory at Delft University of Technology in Nederland has produced several series of yacht models, known together as Delft Systematic Yacht Hull series (DSYHS). This large data base of sailing yacht models today consists of 7 series with a total of approximately 70 models and can be accessed through DSYHS Data Base. According to (Keuning and Katgert, 2008), it is probably the largest series of yacht hulls systematically designed, built, and tested up to now and the series is still expanding. DSYHS is elaborated in much detail in (Keuning and Sonnenberg, 1998). DSYHS has been used for extensive research of sailing yacht hydrodynamics and performance over the past five decades (Keuning and Katgert, 2008; Keuning and Sonnenberg, 1998; Kerwin, 1978; Gerritsma et al., 1981; Gerritsma and Keuning, 1988; Gerritsma et al., 1992; Keuning et al., 1996; Keuning and Binkhorst, 1997). The research presented in this work is based on DSYHS also and deals with prediction of residuary resistance in sailing yachts.

The prediction of total yacht resistance, and particularly its residuary resistance, is very important because of its influence on ship hull design. This prediction should be done with the highest possible accuracy to ensure that the ship operates at optimal speed under most efficient and cost-effective conditions. There are several models for residuary resistance prediction presented in the literature and obtained through regression analysis (Keuning and Katgert, 2008). The variables in these models are different parameters describing hull geometry, and the models are given as sets of polynomials of rather complex structure. Their coefficients are valid only for a specific ship speed, described by a corresponding Froude number. Thus, large look-up tables must be built for each model for different discrete values of Froude number.

This paper proposes and describes several unconventional models for residuary resistance prediction in sailing yachts. The modeling is done on the Yacht Hydrodynamic Data Set available at (Lichman, 2013). The set includes 308 full-scale experiments performed at Delft Ship Hydromechanics Laboratory, which consist of 22 different hull forms. The supplied input parameters are different coefficients concerning the hull geometry and yacht speed: longitudinal position of the center of buoyancy, prismatic coefficient, length to displacement ratio, beam to draught ratio, length to beam ratio, Froude number. The measured variable, i.e. the output, is the residuary resistance per unit weight of displacement.

2 Fuzzy Models of Residuary Resistance

The paper presents two fuzzy models of residuary resistance: a position type and a position – gradient type fuzzy model. The identification of these models is based

on (Sugeno and Yasukawa, 1993). The obtained position type fuzzy model of residuary resistance is of the following form:

$$R^i: IF\ x_3\ is\ A_3^i\ and\ x_5\ is\ A_5^i\ THEN\ y\ is\ B^i; \quad i = 1,2,\cdots,5 \tag{1}$$

and is shown in Figure 1. It has two inputs x_3 and x_5, one output y and five fuzzy rules $R^i (i = 1,2,\cdots, c = 5)$; A_3^i, A_5^i, B^i are fuzzy variables with trapezoidal membership functions, and c is the number of clusters. A deffuzification method known as center of gravity is used to infer the model output $\hat{y}$, and $\hat{y}$ is calculated as the weighted average of the centers of gravity b^i for the consequent membership functions $\mu_{B^i}(y)$, with respect to the weighting factors w^i:

$$\hat{y} = \sum_{i=1}^{c} w^i b^i \Big/ \sum_{i=1}^{c} w^i$$
$$w^i = \min_{1\le j\le n} \mu_{A_j^i}(x_j); 1 \le i \le c \tag{2}$$
$$b^i = \frac{\int y\mu_{B^i}(y)dy}{\int \mu_{B^i}(y)dy}$$

The output of the position type fuzzy model of residuary resistance compared to the actual output is shown in Figure 2. The model performance has been evaluated through the performance index PI, defined as root mean square error (RMSE) of the model output:

$$PI = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{real}^i - y_{model}^i\right)^2} \tag{3}$$

The obtained position-gradient fuzzy model of residuary resistance is shown in Figure 3 and is of the following form:

$$R^i: IF\ x_3\ is\ A_3^i\ and\ x_5\ is\ A_5^i\ THEN\ y\ is\ B^i$$
$$and\ \frac{\partial y}{\partial x_3}\ is\ C_3^i\ and\ \frac{\partial y}{\partial x_5}\ is\ C_5^i; 1 \le i \le c = 5 \tag{4}$$

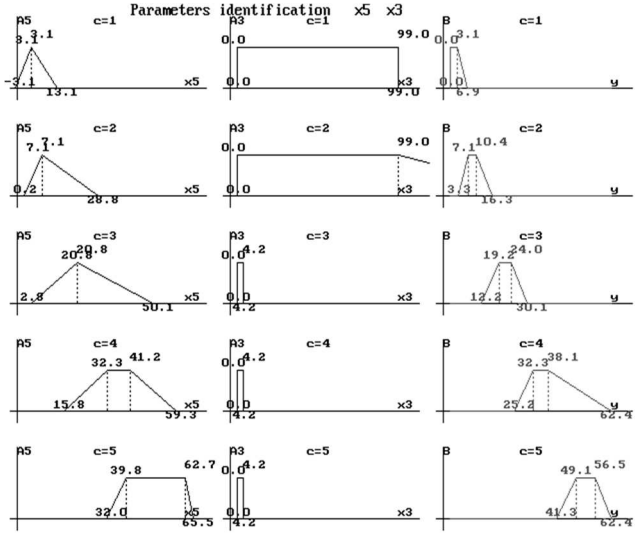


Figure 1. A position type fuzzy model of residuary resistance.

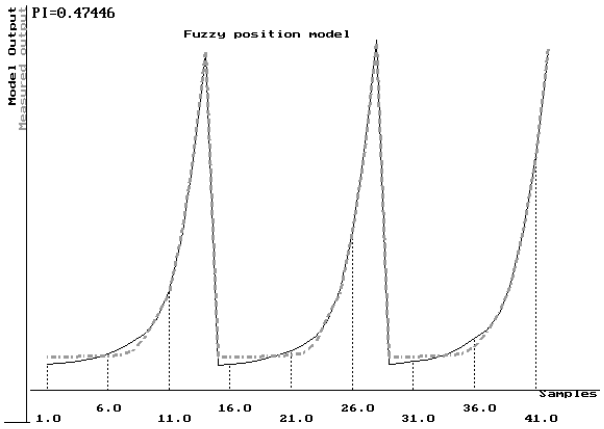


Figure 2. The output of the position type fuzzy model of residuary resistance (dashed line) compared to the actual measured output (solid line).

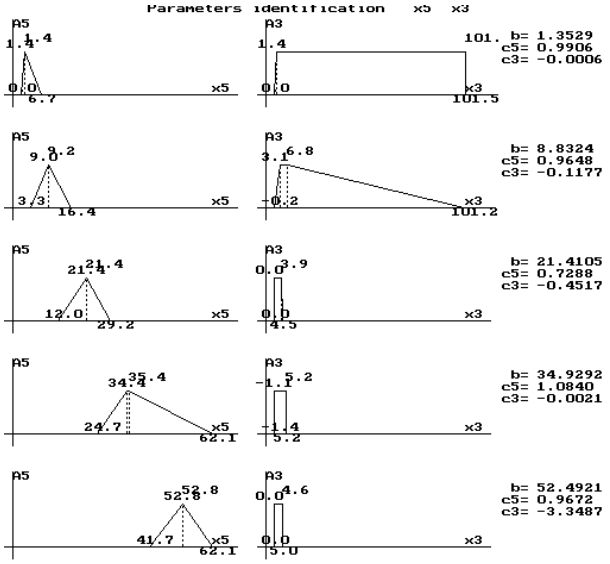


Figure 3. A position-gradient type fuzzy model of residuary resistance.

where A_3^i, A_5^i, B^i, C_3^i, C_5^i are fuzzy variables with trapezoidal membership functions and $\partial y/\partial x_3, \partial y/\partial x_5$ are partial derivatives of the fuzzy rule outputs with respect to the corresponding inputs. The difference between the two fuzzy models is since a position type fuzzy model cannot be built over the whole input space when some data are missing. In these cases, the output of the model can be estimated through extrapolation based on the local fuzzy rules, which leads to a position-gradient fuzzy model. The output of this model is inferred in the following way:

$$\hat{y} = \frac{\sum_{i=1}^{c} w(d^i)\{[b^i + \sum_{j=1}^{n}(d_j^i \times c_j^i)]\}}{\sum_{i=1}^{c} w(d^i)} \tag{5}$$

where b^i and c_j^i are values obtained by defuzzification of $B^i$and C_j^i, respectively; d^i is the distance between the input and the core region of the $i-$th fuzzy rule; d_j^i is a component of d^i on the x_j coordinate axis and $w(d^i) = exp(-d^i)$ is the weight of the $i-$th fuzzy rule with

respect to distance d^i. The performance of the position-gradient type fuzzy model is evaluated through its PI based on RMSE criterion (3) and is shown in Figure 4.

The applied identification algorithm (Sugeno and Yasukawa, 1993) is a very well-known method for fuzzy identification. Several modifications of this identification method have been presented in literature (Tikk, 2002; Haddad, 2008; Kim et al., 1997; Lazarevska and Trpovski, 2000). In this research, the modification presented in (Lazarevska and Trpovski, 2000) is used. The performed algorithm includes parameter identification at each stage of model's rule structure identification process, thus significantly improving the accuracy of the obtained intermediate fuzzy models. As a result, a more efficient and more accurate selection of inputs to the identified fuzzy models is obtained. The parameter identification is done both for the premise and the consequent parameters of the fuzzy rules. In addition, parameter identification is done throughout the process of estimation of partial derivatives of the output with respect to the input variables. As a result, the two obtained fuzzy models very successfully model the given input output data, generating the desired output only by two significant inputs, opposite conventional polynomial models that struggle with many parameters and parameter dependent coefficients. The identification of these models in much more detail is given in (E. Lazarevska, 2016).

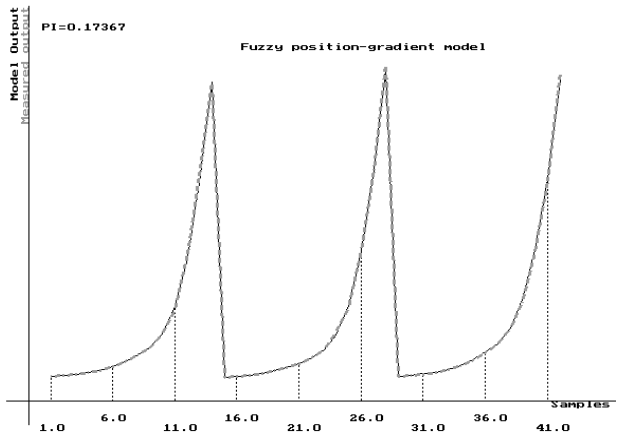


Figure 4. The output of the position-gradient type fuzzy model of residuary resistance (dashed line) compared to the actual measured output (solid line).

3 A Neuro-fuzzy Model for Residuary Resistance Prediction

The neuro-fuzzy model for residuary resistance prediction presented in this paper is based on several excellent papers (Vapnik, 1998; Tipping, 2001, Kim et al., 2006) and is described in detail in (Lazarevska, 2016). The modeling is done on the available input-output data $\{x_k, y_k\}; k = 1,2,\cdots,N$ and the model has the structure of a Takagi and Sugeno (TS) fuzzy model:

$$R^i: IF\ x_1\ is\ A_1^i\ and\ x_2\ is\ A_2^i\ and\ \cdots and\ x_M\ is\ A_M^i\ THEN\ f_i = a_{i1}x_1 + \cdots + a_{iM}x_M + a_{i0}; \quad i = 1,2,\cdots,n \tag{6}$$

where $x_j\ (j = 1,2,\cdots,M)$ are inputs to the fuzzy rules $R^i\ (i = 1,2,\cdots,n)$, A_j^i are appropriate fuzzy sets, a_{ij} are consequent parameters, f_i is the $i-$th local output variable, n is the number of fuzzy rules and M is the dimension of the input data vectors. The fuzzy sets A_j^i are represented by Gaussian type kernel functions:

$$K(x_j, x_{ij}^*) = exp\left[-\frac{(x_j - x_{ij}^*)^2}{2\theta_{ij}^2}\right]; \quad i = 1,2,\cdots,n; j = 1,2,\cdots,M \tag{7}$$

where x_j is the $j-$th feature of the $k-$th input variable x_k, x_{ij}^* is the center and θ_{ij} is the variance of the Gaussian kernel function $K(x_j, x_{ij}^*)$ and $i = 1,,\cdots,n; j = 1,2,\cdots,M$. Thus, the fuzzy IF-THEN rules (6) have the following specific form (Kim et al., 2006):

$$R^i: IF\ x_1\ is\ K(x_1, x_{i1}^*) and\ x_2\ is\ K(x_2, x_{i2}^*) and\ \cdots and\ x_M\ is\ K(x_M, x_{iM}^*)\ THEN\ f_i = a_{i1}x_1 + \cdots + a_{iM}x_M + a_{i0}; \quad i = 1,2,\cdots,n \tag{8}$$

The function $K(x_j, x_{ij}^*)$ in (8) represents the grade of membership of $x_j$with respect to the fuzzy set A_j^i, x_{ij}^* is a relevance vector (RV), θ_{ij} is a kernel parameter, the number of fuzzy rules n equals the number of RVs and $i = 1,,\cdots,n; j = 1,2,\cdots,M$. The number of fuzzy rules and the parameters of the membership functions in (8) are generated automatically by the extended relevance vector learning machine RVM algorithm (Kim et al., 2006). The identification algorithm performs system optimization and generalization simultaneously. The gradient ascent method adjusts the parameters of the kernel functions. The coefficients in the consequent part of the fuzzy rules are determined by the least square method.

The structure of the neuro-fuzzy model of residuary resistance is shown in Figure 5. It is presented as a neural network with six different layers. The first layer is the input layer. It has a total of M nodes, M being the number of elements in the training input vector $\mathbf{x}_\mathrm{k} = (x_{k1}, x_{k2}, \cdots, x_{kM})$. This layer transmits the upcoming input data to the second layer and does not perform any operations over the training input data. The second layer is a fuzzification layer. Each node in this layer has exactly M inputs. The second layer consists of n nodes that represent adequate kernel functions $K(x_j, x_{ij}^*)$. From the fuzzy modeling perspective, the Gaussian kernel function is the membership function $K(x_j, x_{ij}^*) = \mu_{A_j^i}(x_j)$ of the $j-$th fuzzy input x_j with respect to the $i-$th fuzzy rule, its parameters x_{ij}^* and θ_{ij} are premise

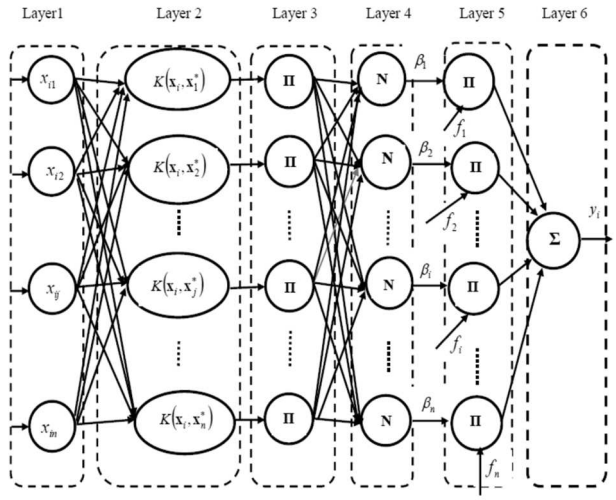


Figure 5. The structure of the neuro-fuzzy model of residuary resistance.

parameters of the corresponding fuzzy rule, M is the number of fuzzy inputs to the neuro-fuzzy model, and the number n of kernel functions is the number of fuzzy rules, i.e. the number of nodes in the second layer. Because of the Gaussian shape of the selected kernel functions, the membership functions of the antecedent part of the fuzzy rules are Gaussian membership functions. From RVM prospective, the center x_{ij}^{*} of the Kernel function is a relevance vector RV, the varianceθ_{ij} is a kernel parameter, and n is the number of RVs. The third layer can be called as the rule layer, since a node in this layer generates the IF part of each fuzzy rule. This layer has n nodes, one for each fuzzy rule, and they compute the firing strength of the associated fuzzy rules using the product of membership functions as T-norm operator. The fourth layer is a normalization layer. It consists of n nodes and each node perform normalization of the firing strength of the associated fuzzy rule. This normalization is done with respect to the sum of the firing strengths of all the fuzzy rules, and the output of each node is the weight β_i of the corresponding fuzzy rule. Each node i in the fifth layer

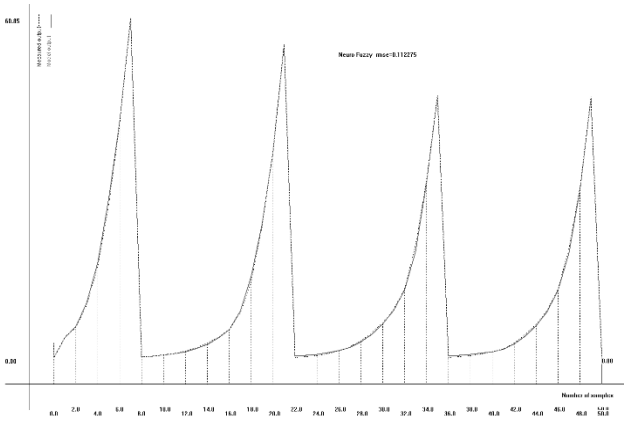


Figure 6. The output of the neuro-fuzzy model of residuary resistance (dashed line) compared to the actual measured output (solid line).

calculates the product of the normalized weight β_i for the $i-$th rule and the local output variable f_i of the fuzzy model. The sixth and the last layer is the output layer. The single node in this layer computes the overall output $f(\mathbf{x})$ of the neuro-fuzzy model as the sum of all incoming signals $\beta_i f_i$ $(i = 1,2,\cdots,n)$. The output of the neuro-fuzzy model for residuary resistance prediction with the obtained relevance vectors, and compared to the actual measured output, is shown in Figure 6.

4 Neural Models of Residuary Resistance based on ELM

Three neural modes of residuary resistance are presented in this section, based on classic and incremental extreme learning machine (ELM). First the classic ELM is used to obtain the desired model of residuary resistance. To perform the modeling, the available experimental data were divided into two sets: training data and testing data, and fixed number of hidden nodes $n = 25$ were assigned for training. Since the approximation performance of classic ELM is generally independent of the type of activation function (Huang and Babri, 1998), the logistic function was chosen for the hidden neurons, and the input parameters of the hidden neurons $\mathbf{w}_\mathrm{i}$ and b_i $(i = 1,,\cdots,n)$ were randomly assigned according to the uniform probability distribution. The neural model of residuary resistance based on the classic ELM is of the following form (Huang et al., 2004):

$$\tilde{y}_k = \sum_{i=1}^{n} \mathbf{v}_i \, g_i(\mathbf{x}_k) = \sum_{i=1}^{n} \mathbf{v}_i \, g(\mathbf{w}_i \mathbf{x}_k + b_i); \quad k = 1,2,\cdots,N \tag{9}$$

where $\mathbf{x}_\mathrm{k} = [x_{k1}\ x_{k2}\ \cdots\ x_{kM}]^T$is the $k-$th input vector of dimension M, $\mathbf{w}_\mathrm{i} = [w_{i1}\ w_{i2}\ \cdots\ w_{iM}]^T$are the weights of the connections between the M input neurons and the $i-$th hidden neuron, $\mathbf{v}_\mathrm{i} = [v_{i1}\ v_{i2}\ \cdots\ v_{iL}]^T$is the vector of the weights defining the connections between the $i-$th hidden neuron and the L output neurons, b_i is the threshold, i.e. the bias of the $i-$th hidden neuron, and $g_i(\mathbf{x}_k)$ is the activation function of the $i-$th hidden neuron; the term $\mathbf{w}_i\mathbf{x}_k$ denotes inner product between $\mathbf{w}_i$and $\mathbf{x}_k$. The hidden neuron output weights can be determined simply and analytically with an adequate least square method yielding the smallest norm least square solution (Huang et al., 2004):

$$\mathbf{V} = \mathbf{W}^{+}\mathbf{Y} \tag{10}$$

where $\mathbf{V}$ is the model adjustable parameter vector, $\mathbf{Y}$ is the model output vector, and $\mathbf{W}^{+}$ stands for the Moore-Penrose generalized inverse of the hidden layer output matrix $\mathbf{W}$ (Rao and Mitra, 1971). However, the conducted research has shown that the approximation performance of classic ELM depends considerably on the values of the arbitrary chosen weights and biases for the hidden layer inputs. Therefore, a simple approach is used in this research to overcome the uncertainty problem with

classic ELM. After the selection of appropriate activation function and the number of hidden neurons, the desired value of the error is set as ε. Then the training of the constructed neural network with ELM is conducted and the error of the obtained model is compared to the preset desired value ε. If the model error is greater than ε, the training process is repeated. Otherwise, it is considered that the obtained model has the desired accuracy. Figure 7 shows the output of the ELM model of residuary resistance obtained in this way compared to the actual output.

The approximation accuracy of ELM certainly depends on the number of hidden layer neurons n, and when n approaches the number of training samples N, the model error approaches zero (Huang and Babri, 1998). However, since too large number of hidden neurons is not a desired neural network feature, (Huang et al., 2006) has proposed a new learning algorithm called as incremental ELM (IELM). The difference between the classic ELM and the IELM is in the addition of new neurons to the hidden layer of the later. The new neurons can be added one at a time, or in groups, and the process of learning continues until the preset maximum number of hidden neurons is reached, or the preset acceptable model error is achieved. As with classic ELM, the input parameters of the hidden layer in IELM are randomly generated and are not adjusted at all during the learning process. When a new hidden neuron is added to the hidden layer, the IELM algorithm does not recalculate the hidden layer output parameters of the existing hidden nodes, and the output weights of the hidden layer are calculated according to (Huang et al., 2006):

$$\boldsymbol{\beta}_n = \mathbf{EH^T}/\mathbf{HH^T} = \sum_{i=1}^{N} e(i)h(i) \Big/ \sum_{i=1}^{N} h^2(i) \qquad (11)$$

where $h(i)$ denotes the activation of the added new hidden node for the i – th training sample, while $e(i)$ is the corresponding residual error before the addition of the new hidden node in question; $\mathbf{H} = [h(1) \quad \cdots \quad h(N)]^{\mathrm{T}}$ is the activation vector of the newly added node for all the training samples, and $\mathbf{E} = [e(1) \cdots e(N)]^T$ is the vector of residual error before the addition of the new hidden neuron. The value of the residual error after the addition of a new hidden neuron is calculated according to (Huang et al., 2006):

$$\mathbf{E} = \mathbf{E} - \boldsymbol{\beta}_\mathbf{n}\mathbf{H}_\mathbf{n} \qquad (12)$$

However, the obtained IELM model does not provide the best possible solution considering the model approximation error, since the output weights of the nodes in the hidden layer are not recalculated after each addition of a new hidden node. To overcome this problem, we have tested a much simpler algorithm than the IELM described above, which is a modification of the classic ELM in a sense that it accepts increasing number of hidden nodes and searches for smallest preset error defined by ε. The algorithm recalculates the output parameters of all the hidden neurons in the hidden layer after every new addition to the hidden layer and performs until the preset maximum number of hidden neurons or the preset desired model error is reached. The performance of our version of IELM residuary resistance model with $n = 30$ and logistic activation function, compared to the actual measured output, is shown in Figure 8. This model has much better performance index than the IELM given by (11) - (12).

To overcome the accuracy issue with IELM, (Huang and Chen, 2007) proposed a modification of IELM called convex incremental ELM (CIELM), which assigns the output hidden layer weights as:

$$\boldsymbol{\beta} = \frac{\mathbf{E}[\mathbf{E} - (\mathbf{Y} - \mathbf{H})]^{\mathbf{T}}}{[\mathbf{E} - (\mathbf{Y} - \mathbf{H})][\mathbf{E} - (\mathbf{Y} - \mathbf{H})]^T} = \frac{\sum_{i=1}^{N} e(i)\{e(i) - [y(i) - h(i)]\}}{\sum_{i=1}^{N}\{e(i) - [y(i) - h(i)]\}^2} \qquad (13)$$

The variables in (13) are defined as in (11). The output of the residuary resistance CIELM model according to (13) is given in Figure 9. It has worse PI than the model in Figure 7, which is due to the fixed random assignment of network input parameters.

The performance indices of the obtained models for residuary resistance prediction are shown in Table 1. The model with our version of IELM has shown the best PI value and the reason for this is found in the fact that

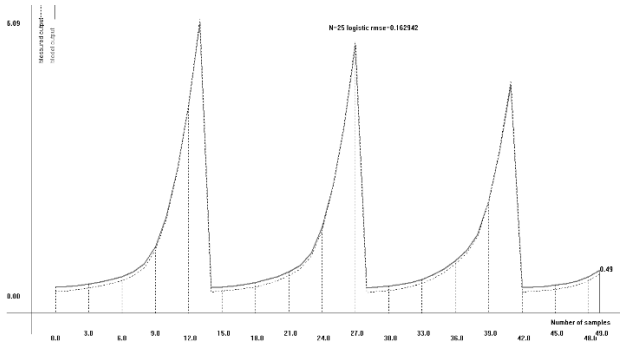


Figure 7. NN model for residuary resistance prediction based on our modification of classic ELM with $\boldsymbol{n = 25}$, logistic activation function and RMSE=0.162942.

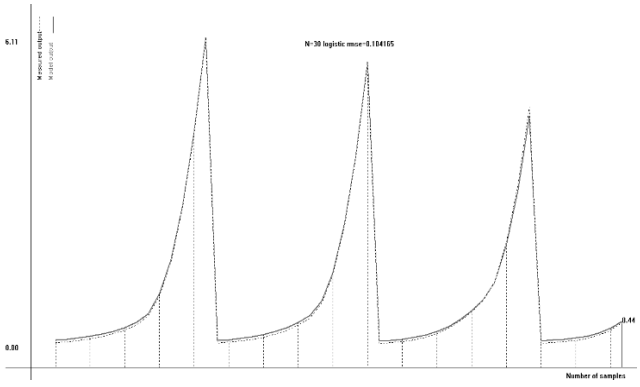


Figure 8. Our version of IELM model for residuary resistance prediction with $\boldsymbol{n = 30}$, logistic activation function and RMSE=0.104165.

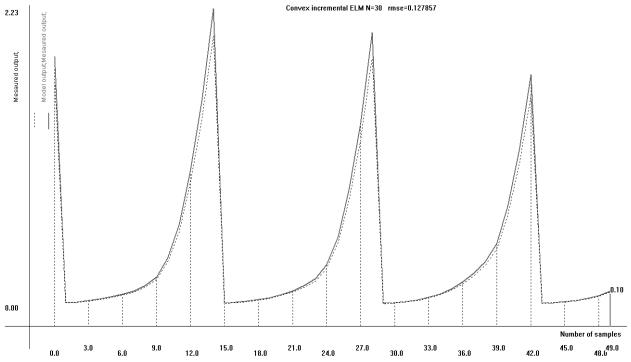


Figure 9. CIELM model for residuary resistance prediction with $\boldsymbol{n = 30}$, logistic activation function and RMSE=0.127857 according to (Huang and Chen, 2007).

whenever a new hidden node is added to the network, all the output parameters of the hidden layer are recalculated, and a search is conducted with different random assignments for the network input parameters until the desired error is reached. IELM and CIELM do not take into consideration the influence of the randomness factor. Once assigned, the input parameters of IELM and CIELM networks are never changed.

Table 1. Comparison of Residuary Resistance Prediction Models obtained by Different Modeling Techniques.

Model	*RMSE*
Position type fuzzy model	0.47446
Position-gradient type fuzzy model	0.17367
Neuro-fuzzy model based on RVM	0.112275
NN model based on our version of ELM	0.162942
NN model based on IELM according to (Huang et al., 2006)	0.143637
NN model based on our version of IELM	0.104165
NN model based on CIELM according to (Huang and Chen, 2007)	0.127857

5 Conclusions

The paper presents six unconventional models for residuary resistance prediction based on fuzzy logic and neural network techniques. All of them possess excellent approximation properties. The applied fuzzy logic approach is especially valuable because of its transparency and ease of interpretability, and because it allows the researcher to determine the most significant input variables that affect the modeled system output and behavior, which is very desirable in cases with large number of input candidates. The position-gradient fuzzy model has a comparable accuracy with the more sophisticated models. The neuro-fuzzy model for residuary resistance prediction has the same excellent approximation property as the rest of the presented models, but it lacks the simplicity and the computational speed of the ELM neural models. The neural models for residuary resistance prediction based on ELM philosophy, have clearly showed that ELM indeed possesses the attributes of extreme simplicity, extremely good approximation performance, and extremely fast computation. Very notably, our version of IELM produces the best approximation performance, meaning the smallest approximation error defined as RMSE, because it takes into consideration the effect of input parameters randomness on the ELM.

References

DSYHS, *DSYHS Data Base*. Available at: http://dsyhs.tudelft.nl/dsyhs.php (accessed 06.02.2016)

J. Gerritsma and J. A. Keuning. Performance of light- and heavy-displacement sailing yachts in waves. In *The 2nd Tampa Bay Sailing Yacht Symp.*, St. Petersburg, 1988.

J.Gerritsma, J. A. Keuning and R. Onnink. *Sailing yacht performance in calm water and in waves*. Report No. 925-P, In *12th Int. Symp. on Yacht Design and Construction HISWA*, 1992.

J. Gerritsma, R. Onnink and A. Versluis. Geometry, resistance and stability of the Delft systematic yacht hull series. *Int. Shipbuilding Progress*, 28: 276– 297, 1981.

A. H. Hadad, T. Gedeon, S, Shahbazi and S. Bahrami. A modified version of Sugeno-Yasukawa modeler. In *13th International CSI Computer Conference*, CSICC 2008 Kish Island, Iran, March 9-11, 2008, 852-856.

G. –B. Huang, and H. A. Babri. Upper bounds on the number of hidden neurons in feedforward networks with arbitrarily bounded nonlinear activation functions. *IEEE Trans. Neural Networks*, 9: 224-229, 1998.

G. –B. Huang, and L. Chen. Convex incremental extreme learning machine. *Neurocomputing*, 70: 3056-3062, 2007.

G. –B. Huang, L. Chen, and C. –K. Siew. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Net.*, 17(4): 879-892, 2006.

G. –B. Huang, Q. –Y. Zhu and C. –K. Siew. Extreme learning machine: A new learning scheme of feedforward neural networks. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN2004)*, 25-29 July, Budapest, Hungary, 2004, 985-990.

J. E. Kerwin. *A velocity prediction program for ocean racing yachts*. Report 78-11, Department of Ocean Engineering, MIT, 1978.

J. A. Keuning and B. J. Binkhorst. Appendage resistance of a sailing yacht hull. In *13th Chesapeake Sailing Yacht Symp.*, 1997.

J. A. Keuning and M. Katgert,. A bare hull resistance prediction method derived from the results of the Delft Systematic Yacht Hull Series extended to higher speeds. In *Int. Conf. on Innovation in High Performance Sailing Yachts*, France, 2008.

J. A. Keuning, R. Onnink, A. Versluis, and A. Van Gulik. The bare hull resistance of the Delft Systematic Yacht Hull Series. In *Int. HISWA Symp. on Yacht Design and Construction*, Amsterdam RAI, 1996.

J. A. Keuning and U. B. Sonnenberg. Approximation of the hydrodynamic forces on a sailing yacht based on the Delft Systematic Yacht Hull Series. In *Int. HISWA Symp. on Yacht Design and Construction*, Amsterdam RAI, 1998, 99-152.

E. Kim, M. Park, S. Ji and M. Park. A new approach to fuzzy modeling. *IEEE Transactions on Fuzzy Systems*, 5(3): 328-337, 1997.

J. Kim, Y. Suga and S. Won. A new approach to fuzzy modeling of nonlinear dynamic systems with noise: relevance vector learning mechanism. *IEEE Trans. on Fuzzy Systems*, 14: 222–231, 2006.

E. Lazarevska. Fuzzy modeling of residuary resistance in sailing yachts. In *XIII International SAUM Conference on Systems, Automatic Control, and Measurements*, Niš, Serbia, November 09th-11th, 2016.

E. Lazarevska. A Neuro-Fuzzy Model of the Residuary Resistance of Sailing Yachts. In *Proceedings of the IEEE Intelligent Systems IS'2016*, Sofia, Bulgaria, 2016, 173-179.

E. Lazarevska and J. Trpovski. A modification of the famous fuzzy model by Sugeno and Yasukawa. In *Proceedings of the International Symposium on Applied Automatic Systems AAS'2000*, Ohrid, Macedonia, 2000, 31-35.

M. Lichman. *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2013.

C. R. Rao and S. K. Mitra. *Generalized Inverse of Matrices and its Applications*. Wiley. 1971.

M. Sugeno and T. Yasukawa. A fuzzy-logic-based approach to qualitative modeling. *IEEE Trans. on Fuzzy Syst.*, 1(1): 7-33, 1993.

D. Tikk, G. Biró, L. T. Kóczy, T. D. Gedeon and K. W. Wong. Improvements and critique on Sugeno's and Yasukawa's qualitative modeling. *IEEE Transactions on Fuzzy Systems*, 10(5): 596-606, 2002.

M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.,* 1: 211–244, 2001.

V. N. Vapnik. *Statistical Learning Theory*. Wiley. 1998.

Flat Patterns Extraction with Collinearity Models

Leon Bobrowski[1,2], Paweł Zabielski[1]

[1]Faculty of Computer Science, Bialystok University of Technology, Bialystok, Poland
[2]Institute of Biocybernetics and Biomedical Engineering, PAS, Warsaw, Poland

l.bobrowski@pb.edu.pl, p.zabielski@pb.edu.pl

Abstract

The term *collinear* (*flat*) pattern means in this article, a set of a large number of feature vectors located on (or near) a plane in multidimensional feature space. Flat patterns extracted from large data set can provide a basis for modeling a local interactions in selected sets of features. Collinear patterns can be discovered in given data set through minimization of some kind of the convex and piecewise linear (*CPL*) criterion functions.

Keywords: data mining, flat patterns, CPL criterion functions, margins

1 Introduction

Data mining tools are used to extraction patterns from multivariate data sets (Hand and Smyth, 2001). The data sets considered in this article are assumed to be formed by the structuralized feature vectors of the same dimensionality and can be represented as the matrices. The word *pattern* means a data subset with a certain type of regularity. The overall goal of the data mining process is to obtain useful information on the basis of the extracted patterns.

The term collinear (*flat*) pattern means a subset of a large number of feature vectors located on and around selected hyperplanes in a certain feature subspace. Discovered collinear patterns can be used also for creating models of linear interaction between many selected features (genes).

Flat patterns can be discovered in data sets through minimization of a certain type of the convex and piecewise linear (*CPL*) criterion functions (Bobrowski, 2014). The basis exchange algorithms can be used for the *CPL* functions minimization. The role the margin in a special type of the *CPL* functions in the flat patterns discovering is examined in the presented paper. A special type of the *CPL* functions gives opportunity to discover the so called *layered patterns* in the feature space.

2 Data subsets in feature subspaces

Let consider the data set C composed of m feature vectors $\mathbf{x}_j = \mathbf{x}_j[n] = [x_{j,1},...,x_{j,n}]^T$ which represent the objects (patients) O_j and belong to a given n-dimensional feature space $F[n]$ ($\mathbf{x}_j \in F[n]$):

$$C = \{\mathbf{x}_j: j = 1,..., m\} \quad (1)$$

The feature space $F[n] = \{x_1,\ldots, x_n\}$ is composed of n features x_i ($i \in I = \{1,...,n\}$). The i-th component $x_{j,i}$ ($x_{j,i} \in R$ or $x_{j,i} \in \{0,1\}$) of the feature vector $\mathbf{x}_j$ is the numerical value of the feature x_i measured on the j-th object O_j.

The k-th feature subspace $F_k[n_k]$ ($F_k[n_k] \subset F[n]$) is made of n_k such features x_i which have the indices i in the subset I_k ($i \in I_k \subset I$) and contains n_k - dimensional reduced vectors $\mathbf{x} = \mathbf{x}[n_k]$ ($\mathbf{x}[n_k] \in F_k[n_k]$). The reduced vectors $\mathbf{x}[n_k]$ are obtained from the feature vectors $\mathbf{x}[n] = [x_1,...,x_n]^T$ by neglecting these components x_i which represent features x_i with the indices i outside the set I_k ($i \notin I_k$). The regular hyperplane $H_k(\mathbf{w}, \theta)$ in the k-th feature subspace $F_k[n_k]$ is defined in the below manner:

$$H_k(\mathbf{w}, \theta) = \{\mathbf{x}: \mathbf{w}^T\mathbf{x} = \theta\} \quad (2)$$

where $\mathbf{x} = [x_1,...,x_{nk}]^T$ is the reduced feature vector ($\mathbf{x} \in F_k[n_k]$), $\mathbf{w} = [w_1,...,w_{nk}]^T$ is the reduced weight vector ($\mathbf{w} \in R^{nk}$) and θ is the threshold ($\theta \in R^1$).

Definition 1: The hyperplane $H_k(\mathbf{w},\theta)$ in the k-th feature subspace $F_k[n_k]$ is *regular* if and if the threshold θ and the weights $w_{,i}$ are different from zero:

$$(\theta \neq 0) \ \textit{and} \ (\forall i \in \{1,\ldots, n_k\}) \quad w_i \neq 0 \quad (3)$$

The k-th data subset $C_k[n_k]$ is constituted by such m_k reduced vectors $\mathbf{x}_j$ ($\mathbf{x}_j \in F_k[n_k]$) which have the indices j from the given subset J_k ($j \in J_k \subset J = \{1,..., m\}$):

$$C_k = C_k[n_k] = \{\mathbf{x}_j: j \in J_k\} \quad (4)$$

The k-th data subset $C_k[n_k]$ (3) can be represented also as the matrix $M[m_k * n_k]$ with the m_k rows and n_k columns. The rows of the matrix $M[m_k * n_k]$ are constituted by particular feature vectors $\mathbf{x}_j$ ($j \in J_k$). Similar representation

of data sets is used in the biclustering methods. We pay attention to the data subsets $C_k[n_k]$ (3) with a *collinear* (*flat*) structure based on regular hyperplanes $H_k(\mathbf{w}, \theta)$ (2) in the feature subspace $F_k[n_k]$.

Definition 2: The data subset $C_k[n_k]$ (4) formed by a large number m_k of reduced vectors $\mathbf{x}_j = \mathbf{x}_j[n_k]$ constitutes the *collinear* (*flat*) *pattern* P_k if all elements $\mathbf{x}_j$ of this subset are located on a regular hyperplane $H_k(\mathbf{w}, \theta)$ (2) in the feature subspace $F_k[n_k]$:

$$(\forall \mathbf{x}_j \in C_k[n_k]) \quad \mathbf{w}^T\mathbf{x}_j = \theta \tag{5}$$

The ε - *layer* $S(\mathbf{w}, \theta)$ in the feature subspace $F_k[n_k]$ is defined on the regular hyperplane $H_k(\mathbf{w}, \theta)$ (2) in the below manner by using a small margin ε ($\varepsilon \geq 0$):

$$S(\mathbf{w}, \theta) = \{\mathbf{x}: \theta - \varepsilon \leq (\mathbf{w} / \| \mathbf{w} \|)^T \mathbf{x} \leq \theta + \varepsilon \tag{6}$$

where $\| \mathbf{w} \| = (\mathbf{w}^T \mathbf{w})^{1/2}$.

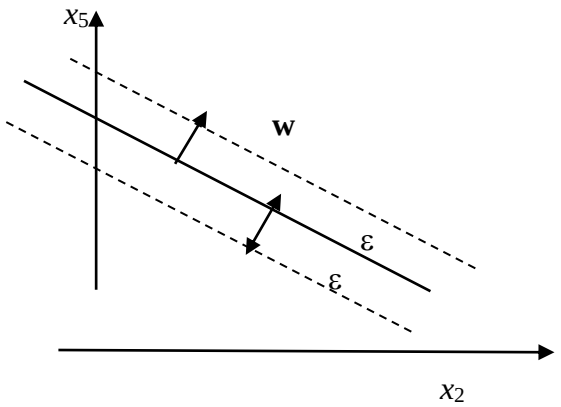


Figure 1. An example of the ε - *layer* $S(\mathbf{w}, \theta)$ (6) in the two-dimensional ($n_k = 2$) feature subspace $F_k = \{x_2, x_5\}$.

Definition 3: The data subset $C_k[n_k]$ (4) has the ε' - *collinear* structure with a margin ε' ($\varepsilon' > 0$) if it exists such weight vector $\mathbf{w}'$ and the threshold θ' that all elements $\mathbf{x}_j$ of this subset are located inside the layer $S(\mathbf{w}', \theta')$ (6):

$$(\forall \mathbf{x}_j \in C_k[n_k]) \quad \theta' - \varepsilon' \leq (\mathbf{w}')^T \mathbf{x}_j \leq \theta' + \varepsilon' \tag{7}$$

where $\| \mathbf{w}' \| = 1$ and $\theta' \neq 0$.

Because the threshold θ' is different from zero ($\theta' \neq 0$) the above inequalities can be given in the following form:

$$(\forall \mathbf{x}_j \in C_k[n_k]) \quad 1 - \varepsilon \leq \mathbf{w}^T \mathbf{x}_j \leq 1 + \varepsilon \tag{8}$$

where $\mathbf{w} = \mathbf{w}' / \theta'$ and $\varepsilon = \varepsilon' / \theta'$.

3 Dual hyperplanes and vertices in the parameter subspaces

Each of reduced feature vector $\mathbf{x}_j$ from the data subset $C_k[n_k]$ (4) defines the below dual hyperplane h_j in the n_k - dimensional parameter subspace R^{nk} ($\mathbf{w} \in R^{nk}$):

$$(\forall \mathbf{x}_j \in C_k[n_k]) \quad h_j = \{\mathbf{w}: \mathbf{x}_j^T \mathbf{w} = 1\} \tag{9}$$

Let consider the set $S_k = \{\mathbf{x}_{j(i)}\}$ of n_k linearly independent reduced feature vector $\mathbf{x}_j[n_k]$ from the subset $C_k[n_k]$ (4)

$$S_k = \{\mathbf{x}_{j(i)}: j(i) \in J_k\} \tag{10}$$

The hyperplanes $h_{j(i)}$ defined by the *basis* vectors $\mathbf{x}_{j(i)}$ from the set S_k (9) intersect at one point (*vertex*) $\mathbf{w}_k$ determined the below equations:

$$(\forall j(i) \in J_k) \quad \mathbf{x}_{j(i)}^T \mathbf{w}_k = 1 \tag{11}$$

The above equations can be given in the matrix form:

$$\mathbf{B}_k^T \mathbf{w}_k = \mathbf{1} \tag{12}$$

where $\mathbf{B}_k = [\mathbf{x}_{j(1)}, \ldots, \mathbf{x}_{j(nk)}]$ is the non-singular matrix called the *k*-th *basis* and $\mathbf{1} = [1,1,\ldots,1]^T$.

The *k*-th vertex $\mathbf{w}_k = [w_{k,1},\ldots,w_{k,nk}]^T$ (11) with the non-zero components $w_{k,i}$ ($w_{k,i} \neq 0$) allows to define the *vertexical hyperplane* $H_k(\mathbf{w}_k, 1)$ in the feature subspace $F_k[n_k]$:

$$H_k(\mathbf{w}_k, 1) = \{\mathbf{x} \in F_k[n_k]: (\mathbf{w}_k)^T \mathbf{x} = 1\} \tag{13}$$

The vertexical hyperplane $H_k(\mathbf{w}_k, 1)$ (12) is defined in the *k*-th feature subspace $F_k[n_k]$ composed from n_k features x_i with the indices *i* belonging to the subset I_k ($i \in I_k$).

Remark 1: All feature vectors $\mathbf{x}_j$ from the subset $C_k[n_k]$ (4) are situated on the hyperplane $H(\mathbf{w},\theta) = \{\mathbf{x}: \mathbf{w}^T\mathbf{x} = \theta\}$ with $\theta \neq 0$, if and only if each vector $\mathbf{x}_j$ defines such dual hyperplane h_j (8) which passes through the vertex $\mathbf{w}_k$ (10).

The *Remark* 1 has been dicussed in the paper.

4 Penalty and criterion functions aimed at extraction of collinear patterns

We consider convex and piecewise linear (*CPL*) penalty functions $\varphi_j(\mathbf{w})$ defined on the n_k - dimensional feature vectors $\mathbf{x}_j$ from the *k*-th data subset $C_k[n_k]$ (4):

$$(\forall \mathbf{x}_j \in C_k[n_k]) \qquad \varphi_j(\mathbf{w}) = \begin{cases} 1 - \varepsilon - \mathbf{w}^T\mathbf{x}_j & \text{if } \mathbf{w}^T\mathbf{x}_j < 1 - \varepsilon \\ 0 & \text{if } 1 - \varepsilon \le \mathbf{w}^T\mathbf{x}_j \le 1 + \varepsilon \\ \mathbf{w}^T\mathbf{x}_j - 1 + \varepsilon & \text{if } \mathbf{w}^T\mathbf{x}_j > 1 + \varepsilon \end{cases} \quad (14)$$

where ε is a small, non-negative parameter (*margin*).

The non-negative function $\varphi_j(\mathbf{w})$ is equal to zero ($\varphi_j(\mathbf{w}) = 0$) if and only if the feature vector $\mathbf{x}_j$ is located in the layer $S(\mathbf{w}, \theta)$ (7) with $\theta = 1$ (Fig. 2)

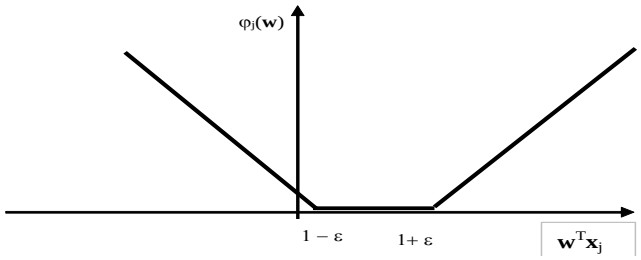


Figure 2. The *j*-th penalty functions $\varphi_j(\mathbf{w})$ (8).

The criterion function $\Phi_k(\mathbf{w})$ is defined as the weighted sum of the penalty functions $\varphi_j(\mathbf{w})$ (8) linked to m_k feature vectors $\mathbf{x}_j$ constituting the subset $C_k \subset C$ (1):

$$\Phi_k(\mathbf{w}) = \Sigma_j \alpha_j \varphi_j(\mathbf{w}) \quad (15)$$

where the positive parameters α_j ($\alpha_j > 0$) are *prices* of particular feature vectors $\mathbf{x}_j$. The parameters α_j may depend on the number m_k of the vectors $\mathbf{x}_j$ in the subset C_k:

$$(\forall \mathbf{x}_j \in C_k) \qquad \alpha_j = 1 / m_k \quad (16)$$

The criterion function $\Phi_k(\mathbf{w})$ (15) is convex and piecewise linear (*CPL*). It can be proved that the minimal value of the function $\Phi_k(\mathbf{w})$ can be found in one of the vertices $\mathbf{w}_k^*$ (11):

$$(\exists\, \mathbf{w}_k^*) \quad (\forall \mathbf{w}) \quad \Phi_k(\mathbf{w}) \ge \Phi_k(\mathbf{w}_k^*) = \Phi_k^* \ge 0 \quad (17)$$

The basis exchange algorithms which are similar to the linear programming allow to find efficiently the optimal vertex $\mathbf{w}_k^*$ (19) constituting the minimal value $\Phi_k(\mathbf{w}_k^*)$ even in the case of large, multidimensional data subsets C_k (4) (Bobrowski, 2014).

For the purpose of the minimization of the criterion function $\Phi_k(\mathbf{w})$ (15) with the penalty functions $\varphi_j(\mathbf{w})$ (14) it is useful to replace each dual hyperplane h_j (9) by the two hyperplanes h_j^+ and h_j^-:

$$(\forall \mathbf{x}_j \in C_k[n_k]) \quad h_j^+ = \{\mathbf{w}: \mathbf{x}_j^T\mathbf{w} = 1 + \varepsilon\} \text{ and } \quad (18)$$

$$h_j^- = \{\mathbf{w}: \mathbf{x}_j^T\mathbf{w} = 1 - \varepsilon\}$$

Theorem 1: If all vectors $\mathbf{x}_j$ from the subset $C_k[n_k]$ (4) can be located inside some ε - *layer* $S(\mathbf{w}',\theta')$ with $\theta' \neq 0$ (5), then the minimal value $\Phi_k(\mathbf{w}_k^*)$ (16) of the criterion function $\Phi_k(\mathbf{w})$ (14) determined on this subset is equal to zero.

Proof: If the reduced vector $\mathbf{x}_j$ is located in the ε-*layer* $S(\mathbf{w}',\theta')$ with $\theta' \neq 0$ (6), then the inequalities (7) are fulfilled for $\mathbf{w} = \mathbf{w}' / \theta'$ and $\varepsilon = \varepsilon' / \theta'$. It means, that the penalty function $\varphi_j(\mathbf{w})$ (14) is equal to zero in the point $\mathbf{w} = \mathbf{w}' / \theta'$. If all elements $\mathbf{x}_j$ of the subset C_k (4) are located inside the *layer* $S(\mathbf{w}',\theta')$, then all the penalty function $\varphi_j(\mathbf{w})$ (13) are equal to zero. It means that the value $\Phi_k(\mathbf{w}_k^*)$ (16) of the criterion function $\Phi_k(\mathbf{w})$ (14) is equal to zero in the point $\mathbf{w} = \mathbf{w}' / \theta'$.

Remark 2: The minimal value $\Phi_k(\mathbf{w}_k^*)$ (17) of the criterion function $\Phi_k(\mathbf{w})$ (15) determined on all elements $\mathbf{x}_j$ of the subset C_k (4) becomes equal to zero for a sufficiently high value of the parameter ε.

For a given data subset $C_k[n_k]$ (4) we can determine the minimum value ε_k of the parameter ε which allows to reset the minimal value $\Phi_k(\mathbf{w}_k^*)$ (17) of the criterion function $\Phi_k(\mathbf{w})$ (15) determined on this subset:

$$\varepsilon_k = min\ \{\varepsilon: \Phi_k(\mathbf{w}_k^*) = 0\} \quad (19)$$

The minimal value ε_k of the parameter ε can be computed for data subset $C_k[n_k]$ (4) through multiple minimization of the criterion function $\Phi_k(\mathbf{w})$ (15) determined on this subset.

Definition 4: The *thickness* ρ_k of the data subset $C_k[n_k]$ (4) is defined to be equal twice the value of the parameter ε_k ($\rho_k = 2\varepsilon_k$) (19).

The minimizing of the criterion function $\Phi_k(\mathbf{w})$ (15) with parameter ε less than ε_k ($0 \le \varepsilon < \varepsilon_k$) allows also to identify in the data subsets $C_k[n_k]$ (4) a part with the greatest collinearity.

5 Vertexical hyperplanes in feature subspaces

The vertexical hyperplane $H_k(\mathbf{w}_k,1)$ (13) in the n_k - dimensional feature subspace $F_k[n_k]$ is defined by using the vertex $\mathbf{w}_k = [w_{k,1},...,w_{k.nk}]^T$ with n_k non-zero components w_i

(4). The vertex $\mathbf{w}_k$ is linked to the k-th basis $\mathbf{B}_k = [\mathbf{x}_{j(1)},\ldots \mathbf{x}_{j(nk)}]$ constituted by n_k linearly independent basis vectors $\mathbf{x}_{j(i)}$ of dimensionality n_k.

The vertexical hyperplane $H_k(\mathbf{w}_k, 1)$ (13) can be represented also in a different manner by using the n_k basis vectors $\mathbf{x}_{j(i)}$ in the feature subspace $F_k[n_k]$:

$$H_k(\mathbf{w}_k, 1) = P_k(\mathbf{x}_{j(1)},\ldots,\mathbf{x}_{j(nk)}) = \\ = \{\mathbf{x}: \mathbf{x} = \alpha_1 \mathbf{x}_{j(1)} + \ldots + \alpha_{nk}\mathbf{x}_{j(nk)}\} \quad (20)$$

where α_i are real numbers ($\alpha_i \in R^1$) which fulfills the below condition:

$$\alpha_1 + \ldots + \alpha_{nk} = 1 \quad (21)$$

Remark 3: The dimensionality of the vertexical hyperplane $P_k(\mathbf{x}_{j(1)},\ldots,\mathbf{x}_{j(nk)})$ (20) is equal to n_{nk} - 1.

Theorem 2: The reduced feature vector $\mathbf{x}_j$ ($\mathbf{x}_j \in F_k[n_k]$) is situated on the *vertexical hyperplane* $P_k(\mathbf{x}_{j(1)}[n],\ldots,\mathbf{x}_{j(rk)}[n])$ (20), where $j(i) \in J_k$ (10) if and only if, the dual hyperplane h_j (9) passes through the vertex $\mathbf{w}_k$ (11) .

The proof of a similar theorem can be found in the paper (Bobrowski, 2014).

Definition 5: The vertexical hyperplane $H_k(\mathbf{w}_k,1)$ (13) supports the *flat pattern* P_k if a large number m_k of the reduced vectors $\mathbf{x}_j$ are located on this hyperplane.

Definition 6: The vertexical hyperplane $H_k(\mathbf{w}_k,1)$ (13) supports the ε - *flat pattern* P_k' if a large number m_k' of the reduced vectors $\mathbf{x}_j$ are located in the ε - *layer* $S(\mathbf{w}_k,1)$ (6) around this hyperplane.

Definition 7: The *rank* r_k of the *flat patterns* P_k or P_k' is equal to the number n_k ($r_k = n_k$) of the basis vectors $\mathbf{x}_{j(i)}$ in the k-th base $\mathbf{B}_k = [\mathbf{x}_{j(1)},\ldots, \mathbf{x}_{j(nk)}]$ (12).

Definition 8: The dimensionality of of the *flat patterns* P_k or P_k' is equal to r_k - 1.

Example 1: The vertexical hyperplane $H_k(\mathbf{w}_k, 1)$ (13) in the feature subspace $F_k[2] = \{x_{i(1)}, x_{i(2)}\}$ represented as the line $l_k(\mathbf{x}_{j(1)},\mathbf{x}_{j(2)})$ spanned (19) by two basis vectors $\mathbf{x}_{j(1)}$ and $\mathbf{x}_{j(2)}$:

$$l_k(\mathbf{x}_{j(1)}, \mathbf{x}_{j(2)}) = \{\mathbf{x}: \mathbf{x} = \alpha \mathbf{x}_{j(1)} + (1 - \alpha) \mathbf{x}_{j(2)}\} \quad (22)$$

where $\alpha \in R^1$.

The *rank* r_k of the *flat patterns* P_k or P_k' supported by the line $l_k(\mathbf{x}_{j(1)},\mathbf{x}_{j(2)})$ (21) is equal 2 ($r_k = 2$).

Example 2: The vertexical hyperplane $H_k(\mathbf{w}_k, 1)$ (13) in the feature subspace $F_k[3] = \{x_{i(1)}, x_{i(2)}, x_{i(3)}\}$ represented as the plane $P_k(\mathbf{x}_{j(1)},\mathbf{x}_{j(2)},\mathbf{x}_{j(3)})$ (19) spanned by three basis vectors $\mathbf{x}_{j(i)}$:

$$P_k(\mathbf{x}_{j(1)},\mathbf{x}_{j(2)},\mathbf{x}_{j(3)}) = \{\mathbf{x}: \mathbf{x} = \alpha_1 \mathbf{x}_{j(1)} + \alpha_2 \mathbf{x}_{j(2)} + \alpha_2 \mathbf{x}_{j(2)}\} \quad (23)$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$ and $\alpha_i \in R^1$.

The *rank* r_k of the *flat patterns* P_k or P_k' supported by the plane $P_k(\mathbf{x}_{j(1)},\mathbf{x}_{j(2)},\mathbf{x}_{j(3)})$ (22) is equal 3 ($r_k = 3$).

The *flat patterns* P_k or P_k' can be extracted from the data set C (1) trough minimization of the criterion functions $\Phi_k(\mathbf{w})$ (15).

6 Properies of the criterion functions $\Phi_k(\mathbf{w})$

The criterion function $\Phi(\mathbf{w})$ is defined as the weighted sum (15) of the penalty functions $\varphi_j(\mathbf{w})$ (8) on the basis of m feature vectors $\mathbf{x}_j = \mathbf{x}_j[n]$ ($\mathbf{x}_j \in F[n]$) constituting the data set C (1). The criterion function $\Phi_k(\mathbf{w})$ is defined (15) on the basis of m_k reduced vectors $\mathbf{x}_j$ ($\mathbf{x}_j \in F_k[n_k]$) from the subset $C_k[n_k]$ (4). In accordance with the *Definition* 1, the data subset $C_k[n_k]$ (4) is the k-th the *flat pattern* P_k if all elements $\mathbf{x}_j$ of this subset can be located on the hyperplane $H_k(\mathbf{w}_k,1)$ (12). We can infer from the *Theorem* 2, that a large number m_k of the vectors $\mathbf{x}_j$ located on the vertexical hyperplane $H_k(\mathbf{w}_k,1)$ (12) causes passing m_k dual hyperplanes h_j (9) through the vertex $\mathbf{w}_k$ (11). In result, the vertex $\mathbf{w}_k$ (11) becomes highly degenerated. The minimization of the criterion functions $\Phi_k(\mathbf{w})$ (15) allows to discover highly degenerated vertices $\mathbf{w}_k$ (11) and, in result, to extract flat patterns P_k.

The folowing properies of the criterion functions $\Phi_k(\mathbf{w})$ (15). can be useful in flat patterns extraction from the data set C (1). The minimal value $\Phi_k(\mathbf{w}_k^*)$ (18) of the criterion function $\Phi_k(\mathbf{w})$ (14) can be characterized by two below *monotonocity properties* (Bobrowski, 2014):

i. The positive monotonocity due to reduction of feature vectors $\mathbf{x}_j$

Neglecting some feature vectors $\mathbf{x}_j$ the data set C (1) cannot result in an increase of the minimal value $\Phi_k(\mathbf{w}_k^*)$ (17) of the criterion function $\Phi_k(\mathbf{w})$ (15):

$$(C_{k'} \subset C_k) \Rightarrow (\Phi_{k'}^* \leq \Phi_k^*) \quad (24)$$

where the symbol Φ_k^* stands for the minimal value (18) of the criterion function $\Phi_k(\mathbf{w})$ (14) defined on the elements $\mathbf{x}_j$ of the subset C_k ($\mathbf{x}_j \in C_k$).

The implication (22) can be proved by the fact that omission of certain feature vectors $\mathbf{x}_j$ results in omission of certain non-negative components $\alpha_j \varphi_j(\mathbf{w})$ (14) in the criterion function $\Phi_k(\mathbf{w})$ (15).

ii. The negative monotonicity due to reduction of features x_i

The reduction of the feature space $F_k[n_k]$ to $F_{k'}[n_{k'}]$ by neglecting some features x_i cannot result in a decrease of the minimal value $\Phi_k(\mathbf{w}_k^*)$ (17) of the criterion function $\Phi_k(\mathbf{w})$ (15):

$$(F_{k'}[n_{k'}] \subset F_k[n_k]) \Rightarrow (\Phi_{k'}^* \geq \Phi_k^*) \tag{25}$$

where the symbol Φ_k^* stands for the minimal value (17) of the criterion function $\Phi_k(\mathbf{w})$ (15) defined on the reduced vectors $\mathbf{x}_j'$ ($\mathbf{x}_j' \in F_{k'}[n_{k'}]$, $n_{k'} < n_k$). The implication (25) results from the fact that the omission of certain features x_i is equivalent to imposing an additional constraint "$w_i = 0$" during the minimization (17) in the parameter space R^{nk}.

Theorem 3: The minimal value $\Phi_k(\mathbf{w}_k^*)$ (17) of the criterion function $\Phi_k(\mathbf{w})$ (15) defined on reduced feature vectors $\mathbf{x}_j$ from the subset C_k (4) does not depend on linear, non-singular transformations of the feature vectors $\mathbf{x}_j$ from this subset:

$$\Phi_k'(\mathbf{w}_k') = \Phi_k(\mathbf{w}_k^*) \tag{26}$$

where $\Phi_k'(\mathbf{w}_k')$ is the minimal value of the criterion functions $\Phi_k'(\mathbf{w})$ (15) defined on the transformed feature vectors $\mathbf{x}_j'[n]$:

$$(\forall \mathbf{x}_j \in C_k) \quad \mathbf{x}_j' = \boldsymbol{A}\, \mathbf{x}_j \tag{27}$$

where $\boldsymbol{A}$ is a non-singular matrix of dimension (n_k x n_k) ($\boldsymbol{A}^{-1}$ exists).

Proof: The values $\varphi_j'(\mathbf{w}[n])$ of the penalty function $\varphi_j(\mathbf{w}[n])$ (15) in a point $\mathbf{w}'[n]$ are defined in the below manner on the transformed feature vectors $\mathbf{x}_j'[n]$ (26):

$$(\forall \mathbf{x}_j \in C_k) \quad \varphi_j'(\mathbf{w}') = |1 - (\mathbf{w}')^T \mathbf{x}_j'| = |1 - (\mathbf{w}')^T \boldsymbol{A} \mathbf{x}_j| \tag{28}$$

If we take (17)

$$\mathbf{w}' = (\boldsymbol{A}^T)^{-1}\, \mathbf{w}_k^* \tag{29}$$

we obtain the below result

$$(\forall \mathbf{x}_j \in C_k) \quad \varphi_j'(\mathbf{w}') = \varphi_j(\mathbf{w}_k^*) \tag{30}$$

The above equation mean that the value $\Phi_k'(\mathbf{w}')$ of the criterion functions $\Phi_k'(\mathbf{w})$ (15) defined in the point $\mathbf{w}'$ (29) on the transformed feature vectors $\mathbf{x}_j'$ (26) is equal to the minimal value $\Phi_k(\mathbf{w}_k^*)$ (17) of the criterion function $\Phi_k(\mathbf{w})$ (15) defined on the feature vectors $\mathbf{x}_j$ ($\mathbf{x}_j \in C_k[n_k]$ (4)).

7 Procedure of flat pattrerns extraction

The *collinear* (*flat*) patterns P_k (*Def.* 2) can be extracted from the data set C (1) through multiple minimization of the criterion functions $\Phi_k(\mathbf{w})$ (15). The procedure *Vertex* can be used for this purpose (Bobrowski, 2014). The basic form of this procedure is given below with using the counter l:

Procedure Vertex

i. $l = 1$; $C_l = C$ (1); (31)

ii. Define the criterion function $\Phi_l(\mathbf{w})$ (15) on all elements $\mathbf{x}_j$ of the data set C_l and find the optimal vertex $\mathbf{w}_l^*$ (11) which constitutes the minimal value $\Phi_k(\mathbf{w}_l^*)$ (17) of this function.

iii. If $\Phi_l(\mathbf{w}_l^*) = 0$, then the procedure is **stopped** in the optimal vertex $\mathbf{w}_l^*$, otherwise the next step is executed

iv. Find the vector $\mathbf{x}_{j'}$ in the feature subset C_l with the highest value of the penalty function $\varphi_j(\mathbf{w})$ (14) in the optimal vertex $\mathbf{w}_l^*$ (18):

$$(\forall \mathbf{x}_j \in C_l) \quad \varphi_{j'}(\mathbf{w}_l^*) \geq \varphi_j(\mathbf{w}_l^*) \tag{32}$$

or with an additional emphasis on the parameters α_j (15):

$$(\forall \mathbf{x}_j \in C_l) \quad \alpha_{j'}\, \varphi_{j'}(\mathbf{w}_l^*) \geq \alpha_j\, \varphi_j(\mathbf{w}_l^*) \tag{33}$$

v. Remove the feature vector $\mathbf{x}_{j'}$ from the subset C_l:

$$C_l \rightarrow C_l / \mathbf{x}_{j'} \tag{34}$$

vi. Increase the counter k:

$$l \rightarrow l + 1 \tag{35}$$

vii. Go to the step *ii.*

The resulting set C_k^* (4) of feature vectors $\mathbf{x}_j$, the set J_k^* (4) of these vectors indices j, and the optimal vertex $\mathbf{w}_k^*$ (11) can be created as a result of the *Vertex* procedure:

$$C_k^* = C_l \ (34) = \{\mathbf{x}_j : j \in J_k^*\} \quad \textit{and} \quad \mathbf{w}_k^* = \mathbf{w}_l^* \tag{36}$$

It can be proved that the *Procedure Vertex* is **stopped** in some vertex $\mathbf{w}_k^*$ (17) of after finite number of steps l. The vertex $\mathbf{w}_k^*$ resulting from the procedure fulfils the below condition (17):

$$(\forall \mathbf{w}) \quad \Phi_k(\mathbf{w}) \geq \Phi_k(\mathbf{w}_k^*) = 0 \tag{37}$$

where the criterion function $\Phi_k(\mathbf{w})$ (15) is determined on all elements $\mathbf{x}_j$ of such reduced data subset C_k which results from the *Vertex* procedure.

The vertex $\mathbf{w}_k^* = [w_{k,1}^*, \ldots, w_{k,nk}^*]^T$ (36) obtained from the procedure *Vertex* (31) should be *regularized* before

using it in the definition of the *vertexical hyperplane* $H_k(\mathbf{w}_k^*, 1)$ (13). The regularization process means in this case the neglecting of such components $w_{k,i}^*$ in the vector $\mathbf{w}_k^*$ which are equal to zero $(w_i = 0)$ (*Def.* 1). The regularization means additionally the neglecting of such features x_i and components $x_{j,i}$ of the feature vectors $\mathbf{x}_j = [x_{j,1},...,x_{j,n}]^T$ from the reduced data subset C_k which are linked to weights $w_{k,i}^*$ equal to zero $(w_{k,i}^* = 0)$:

$$(\forall i \in \{1,...,n\}) \ (\forall j \in J_k(36)) \quad (38)$$
if $(w_{k,i}^* = 0)$, ***then*** (the *i*-th feature x_i and the *i*-th component $x_{j,i}$ of the *j*-th feature vector $\mathbf{x}_j$ are neglected)

Remark 4: The reduction of feature vectors $\mathbf{x}_j$ in the set C (1) in accordance with the procedure *Vertex* combined with the reduction of features x_i in accordance with the rule (38) leads in a finite number of steps l to the extraction of the collinear data subset $C_k[n_k]$ (5) composed of m_k reduced vectors $\mathbf{x}_j$ $(\mathbf{x}_j \in F_k[n_k])$ which fulfill the equation (5).

The *Remark* 4 can justify directly on the basis of the description of the procedure *Vertex* and the rule (38).

Remark 5: If the number m_k of elements $\mathbf{x}_j$ of the final subset $C_k[n_k]$ obtained in result of the procedure *Vertex* and the rule (38) is a large enough, than this subset constitutes the flat pattern P_k (5) (*Def.* 2).

The *Procedure Vertex* (31) gives possibility for discovering and extraction more than one flat pattern P_k (5) from a given data set C (1). For this purpose the data set data set C (1) can be reduced in subsequent cycles k of the below procedure:

During the first cycle ($k = 1$), the *Procedure Vertex* (31) is activated on the data set C_1 equal to the full data set C (1) and ends with the set C_1^* (36).

The initial data set $C_1 = C$ (1) is reduced by the final set C_1^* (36) after the first cycle:

$$C_2 = C_1 \ / C_1^* = C \ / C_1^* \quad (39)$$

The second cycle ($k = 2$) is activated on the data set C_2 and ends with the set C_2^*:

$$C_3 = C_2 \ / C_2^* \quad (40)$$

The third cycle ($k = 3$) is activated on the set C_3 and so on.

The above procedure should be stopped after extraction of an adequate number K of the flat patterns P_k (5). The stop criterion should take into account that the numbers m_k of elements $\mathbf{x}_j$ in the final subsets C_k^* (36) can not be too small.

8 Examples of experimental results

The computational pro/cedures described in this paper are currently being implemented. The first results of the calculations are shown in this paragraph.

Two synthetic data sets D_1 and D_2, has been created for the purpose of the computational experiments. The set D_1 contained $m_1 = 100$ two-dimensional feature vectors $\mathbf{x}_j$ $(\mathbf{x}_j \in R^2)$. The set D_2 contained $m_2 = 100$ three-dimensional feature vectors $\mathbf{x}_j$ $(\mathbf{x}_j \in R^3)$. The data sets D_1 and D_2 were ***collinear***. It means in this case, that elements $\mathbf{x}_j$ of each set D_k ($k = 1, 2$) has been located on the vertexical line $l_k(\mathbf{x}_{j(1)}, \mathbf{x}_{j(2)})$ (22) defined by two basic feature vectors $\mathbf{x}_{j(1)}$ and $\mathbf{x}_{j(2)}$ contained in the basis $\mathbf{B}_k$ (12):

The basic feature vectors $\mathbf{x}_{j(1)}$ and $\mathbf{x}_{j(2)}$ (25) were pre-selected as:

$$P_1: \ \mathbf{x}_{j(1)} = [1,0]^T \ \textit{and} \ \mathbf{x}_{j(2)} = [0,1]^T \ \textit{and} \quad (41)$$
$$P_2: \ \mathbf{x}_{j(1)} = [1,1,0]^T \ \textit{and} \ \mathbf{x}_{j(2)} = [0,1,1]^T$$

The computational experiments were carried out both on the collinear data set P_1 with added *outliers*. as well as on the set P_2 without *outliers*. The term *outliers* means here such additional feature vectors $\mathbf{x}_j$ which were not located on the vertexical line $l_k(\mathbf{x}_{j(1)}, \mathbf{x}_{j(2)})$ (22). The outlier feature vectors $\mathbf{x}_j$ were generated in accordance with the normal distribution $N_2(\mathbf{0}, \boldsymbol{I})$ with the unit covariance matrix $\boldsymbol{I}$.

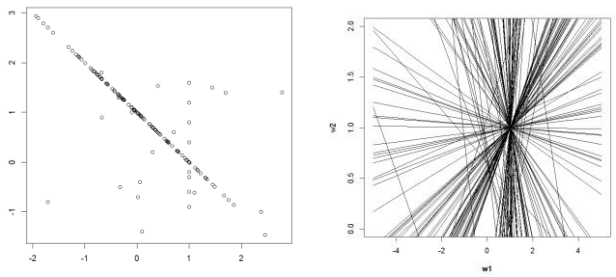


Figure 3. Representations of the collinear pattern P_1 with added outliers in the two-dimensional feature space (*left*) and in the parameter space (*right*).

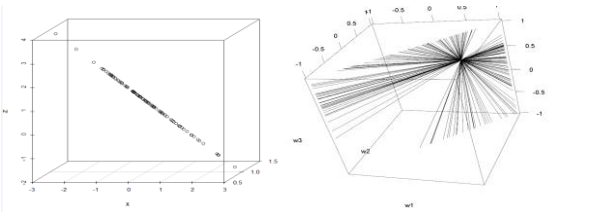


Figure 4. Representations of the collinear pattern P_2 (41) without outliers in the three-dimensional feature space (*left*) and in the parameter space (*right*).

The computational experiments allowed to extract the flat patterns P_1 and P_1 (41) from the data sets given in the feature space.

9 Concluding remarks

Collinear patterns P_k (*Def.* 2) can be discovered in large, high-dimensional data sets C (1) through minimization of the convex and piecewise linear (*CPL*) criterion functions $\Phi_k(\mathbf{w})$ (12).

Discovering collinear patterns P_k can be linked to a search for degenerated vertices (9) in the parameter space.

The proposed by us method of discovering collinear patterns on the basis of the CPL functions can be compared with the methods based on the Hough transformation used in computer vision for detection lines and curves in pictures (Duda and Hart, 1972; Ballard, 1981).

Acknowledgments: The present study was supported by a grant S/WI/2/2013 from Bialystok University of Technology and founded from the resources for research by Ministry of Science and Higher Education.

References

D. H. Ballard. Generalizing the *Hough Transform* to Detect Arbitrary Shapes, *Pattern Recognition*, 13(2):111-122, 1981

L. Bobrowski. Discovering main vertexical planes in a multivariate data space by using *CPL* functions, *ICDM* 2014, Ed. Perner P., Springer Verlag, Berlin 2014

O. R. Duda and P. E. Hart. Use of the Hough Transformation to Detect Lines and Curves in Pictures, *Communications of Association for Computing Machinery*, 15(1):11-15, 1972.

D. Hand, P. Smyth, and H. Mannila. *Principles of data mining*, MIT Press, Cambridge (2001)

Simulating the Effect of Adaptivity on Randomization

Adam Viktorin Roman Senkerik Michal Pluhacek

Faculty of Applied Informatics, Tomas Bata University in Zlin, T. G. Masaryka 5555, 760 01 Zlin, Czech Republic,
{aviktorin,senkerik,pluhacek}@fai.utb.cz

Abstract

This paper compares the development of multi-chaotic system during the optimization process on three classical benchmark functions – Rosenbrock, Rastrigin and Ackley. The multi-chaotic system involves five different randomizations based on discrete chaotic maps (Burgers, Delayed Logistic, Dissipative, Lozi and Tinkerbell) and the probability of their selection is adjusted according to the development of the optimization task. Two variants of Differential Evolution (DE) are used in order to simulate the effect of adaptivity on the randomization probability adjustment process. First non-adaptive variant is DE with rand/1 mutation strategy and the second adaptive variant is novel Success-History based Adaptive DE (SHADE).

Keywords: randomization, differential evolution, SHADE, chaos, parent selection

1 Introduction

The Differential Evolution (DE) has played a significant role in optimization and outperformed other Evolutionary Computation Techniques (ECT) in many cases (Price *et al*, 2006; Kim *et al*, 2007; Chauhan *et al*, 2009; Babu and Jehan, 2003). The original version was introduced in 1995 (Storn and Price, 1995) and since then has been thoroughly studied and improved. One branch of improvement is in adapting its control parameters to the solved optimization task. The examples of adaptive variants are jDE (Brest *et al*, 2006), JADE (Zhang and Sanderson, 2009) and Success-History based Adaptive DE (SHADE) (Tanabe and Fukunaga, 2013). The last listed is used as a representative of adaptive DE variants in this research paper.

One of the recent research directions in ECT is the studying of effect of different randomizations on various parts of the evolutionary algorithms and swarm intelligence algorithms. Especially, the chaotic maps are often used as Pseudo-Random Number Generators (PRNGs) instead of the classical ones with uniform distribution (dos Santos Coelho *et al*, 2014; Senkerik *et al*, 2015b; Caponeto *et al*, 2003) or combinations of multiple chaotic systems with some sort of switching mechanism (Pluhacek *et al*, 2014; Senkerik *et al*, 2015a).

The main research question of this paper is whether there is a randomization or their combination, that would be preferred in parent selection process of non-adaptive and adaptive variants of DE and if the preferences vary for these two variants. In order to simulate that, the multi-chaotic framework containing five different chaotic map based PRNGs was created and probability adjustment process, which mirrors the preference is presented. DE and SHADE algorithms with multi-chaotic framework are tested on three classic benchmark functions – Rosenbrock, Rastrigin and Ackley and the resulting probability development is reported.

The remainder of this paper is structured as follows. Section 2 illustrates chaotic maps and their use as a PRNGs. Section 3 describes DE, SHADE and multi-chaotic framework with pseudo-codes. Following Section 4 is devoted to experiments and results and the whole paper is concluded in Section 5.

2 Chaotic Maps

The chaotic maps are systems generated continuously from a single initial position by simple equations. The current coordinates are generated from the previous ones, consequently creating a system which is extremely dependent on the initial position. The generated chaotic sequence varies for different initial positions. Therefore, the generation of the initial position is randomized to obtain unique chaotic sequences. The generation of starting positions is carried out by PRNG with uniform distribution. Chaotic map equations may also contain control parameters, which determine the chaotic behavior and dynamics.

Chaotic systems used in this research, with their generating equations, control parameter values and initial position generator settings are depicted in Table 1. All the control parameter values were set according to previous experiments and suggestions in literature (Sprott and Sprott, 2003).

The multi-chaotic framework used for parent selection presented in this paper uses five chaotic maps – Burgers, Delayed Logistic, Dissipative, Lozi and Tinkerbell. Each of the chaotic map based PRNGs has different probability distribution and unique sequencing, which may be beneficial for the parent selection process where the obtained parent vector combinations exhibit a different dynamic than that of

parent vector combinations selected by a PRNG with uniform distribution. The distribution of 5,000 real numbers from range [0, 1] generated by each chaotic map can be seen in Figure 1 and Figure 2. It is important to mention the differences between these chaotic systems. While Lozi and Dissipative chaotic maps tend to generate values from the whole range without any clear preference, other three chaotic maps visibly favor values close to the left end of the specified range. In fact, the distribution identifier in Wolfram Mathematica 10.2 software shown that the distribution of Lozi map generated values is beta distribution with shape parameters $\alpha = 1.03457$ and $\beta = 1.56153$ and similarly the distribution of Dissipative map generated values was identified as uniform with range [0.00034, 0.99857]. Other three were identified as mixtures of different distributions.

Table 1. Chaotic maps, generating equations, control parameters and initial position ranges

Chaotic map	Equations	Parameters	Initial position
Burgers	$X_{n+1} = aX_n - Y_n^2$ $Y_{n+1} = bY_n + X_nY_n$	$a = 0.75$ $b = 1.75$	$X_0 = (-0.1, -0.01)$ $Y_0 = (0.01, 0.1)$
Delayed Logistic	$X_{n+1} = AX_n(1 - Y_n)$ $Y_{n+1} = X_n$	$A = 2.27$	$X_0 = Y_0 = (0.8, 0.9)$
Dissipative	$X_{n+1} = X_n + Y_{n+1} \pmod{2\pi}$ $Y_{n+1} = bY_n + k\ sinX_n \pmod{2\pi}$	$b = 0.1$ $k = 8.8$	$X_0 = Y_0 = (0, 0.1)$
Lozi	$X_{n+1} = 1 - a\lvert X_n\rvert - bY_n$ $Y_{n+1} = X_n$	$a = 1.7$ $b = 0.5$	$X_0 = Y_0 = (0, 0.1)$
Tinkerbell	$X_{n+1} = X_n + Y_n + aX_n + bY_n$ $Y_{n+1} = 2X_nY_n + cX_n + dY_n$	$a = 0.9$ $b = -0.6$ $c = 2$ $d = 0.5$	$X_0 = (-0.1, -0.01)$ $Y_0 = (0, 0.1)$

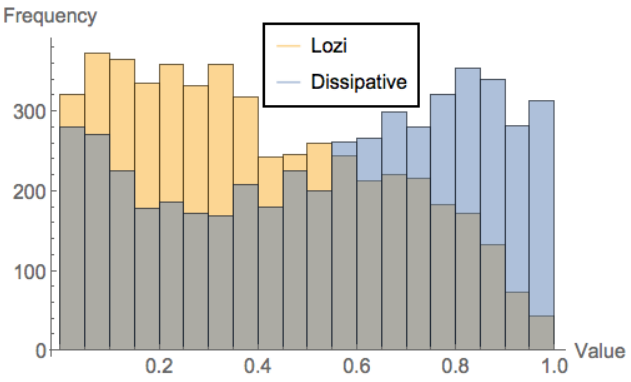


Figure 1. Lozi and Dissipative generated values in histogram.

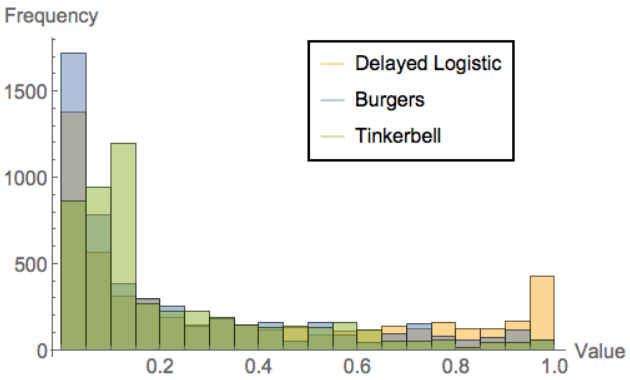


Figure 2. Delayed Logistic, Burgers and Tinkerbell generated values in histogram.

In order to use chaotic maps as PRNGs, the transformation rule had to be developed. The process of obtaining the i-th random integer value $rndInt_i$ from the chaotic map is presented in (1).

$$rndInt_i = \text{round}\left(\frac{\text{abs}(X_i)}{\max(\text{abs}(X_{i\in N}))} * (maxRndInt - 1)\right) + 1 \quad (1)$$

Where $\text{abs}(X_i)$ is the absolute value of the i-th generated X coordinate from the chaotic sequence of

length N, max(abs($X_{i \in N}$)) is a maximum value of all absolute values of generated X coordinates in chaotic sequence. The function round() is common rounding function and *maxRndInt* is a constant to ensure that integers will be generated in the range [1, *maxRndInt*].

3 Differential Evolution, Success-History based Adaptive Differential Evolution and Multi-Chaotic Framework

This section describes DE and SHADE algorithms and their individual parts. It also covers the proposed multi-chaotic framework for parent selection.

3.1 Differential Evolution and Success-History based Adaptive Differential Evolution

The DE (Storn and Price, 1995) has four static control parameters – number of generations G_{max}, population size NP, scaling factor F and crossover rate CR. In the evolutionary process of DE, these four parameters remain unchanged and depend on the user initial setting. SHADE algorithm, on the other hand, adapts the F and CR parameters during the evolution. The values that brought improvement to the optimization task are stored into according historical memories M_F and M_{CR}.

The whole process can be divided into five parts – initialization, mutation strategy with parent selection, crossover, elitism and historical memory update (the last one is only in SHADE algorithm).

3.1.1 Initialization

DE: The initial population of size NP is generated randomly from the objective space. Control parameters F, CR and G_{max} are set.

SHADE: The initial population is generated as in DE, external archive of inferior solutions A is initialized empty and has a maximum size of NP. Both historical memories have the same size H and are initialized to $M_{CR,i} = M_{F,i} = 0.5$ for ($i = 1, \ldots, H$).

3.1.2 Mutation Strategy with Parent Selection

DE: The selected mutation strategy for DE algorithm in this paper is rand/1 (Storn and Price, 1995), which combines 3 different randomly selected parent vectors $\boldsymbol{x}_{r1,G}$, $\boldsymbol{x}_{r2,G}$ and $\boldsymbol{x}_{r3,G}$ from current generation G. Additionally, parent vectors have to differ from the original vector $\boldsymbol{x}_{i,G}$, therefore $\boldsymbol{x}_{i,G} \neq \boldsymbol{x}_{r1,G} \neq \boldsymbol{x}_{r2,G} \neq \boldsymbol{x}_{r3,G}$. All three parents are selected by the PRNG with uniform distribution. The rand/1 mutation is depicted in (2) where $\boldsymbol{v}_{i,G}$ is the resulting mutated vector and F is the static scaling factor.

$$\boldsymbol{v}_{i,G} = \boldsymbol{x}_{r1,G} + F\left(\boldsymbol{x}_{r2,G} - \boldsymbol{x}_{r3,G}\right) \quad (2)$$

SHADE: In the original version of SHADE algorithm (Tanabe andY Fukunaga, 2013), parent selection for mutation strategy is carried out by the PRNG with uniform distribution. The mutation strategy used in SHADE is current-to-*p*best/1 and uses four parent vectors – current i-th vector $\boldsymbol{x}_{i,G}$, vector $\boldsymbol{x}_{pbest,G}$ randomly selected from the $NP \times p$ best vectors (in terms of objective function value) from current generation G. The p value is randomly generated by uniform PRNG $U[p_{min}, 0.2]$, where $p_{min} = 2/NP$. Third parent vector $\boldsymbol{x}_{r1,G}$ is randomly selected from the current generation and last parent vector $\boldsymbol{x}_{r2,G}$ is also randomly selected, but from the union of current generation G and external archive A. Also, vectors $\boldsymbol{x}_{pbest,G}$, $\boldsymbol{x}_{i,G}$, $\boldsymbol{x}_{r1,G}$ and $\boldsymbol{x}_{r2,G}$ has to differ, $\boldsymbol{x}_{pbest,G} \neq \boldsymbol{x}_{i,G} \neq \boldsymbol{x}_{r1,G} \neq \boldsymbol{x}_{r2,G}$. The mutated vector $\boldsymbol{v}_{i,G}$ is generated by (3).

$$\boldsymbol{v}_{i,G} = \boldsymbol{x}_{i,G} + F_i\left(\boldsymbol{x}_{pbest,G} - \boldsymbol{x}_{i,G}\right) + F_i\left(\boldsymbol{x}_{r1,G} - \boldsymbol{x}_{r2,G}\right) \quad (3)$$

The i-th scaling factor F_i is generated from a Cauchy distribution with the location parameter $M_{F,r}$ (selected randomly from the scaling factor historical memory M_F) and scale parameter value of 0.1 (4). If $F_i > 1$, it is truncated to 1 also if $F_i \leq 0$, (4) is repeated.

$$F_i = C\left[M_{F,r}, 0.1\right] \quad (4)$$

DE and SHADE: If any of the features of the mutated vector $\boldsymbol{v}_{i,G}$ is outside the boundaries of objective space in that dimension [$x_{j,min}$, $x_{j,max}$], it is constrained as shown in (5).

$$v_{j,i,G} = \begin{cases} \left(x_{j,min} + x_{j,i,G}\right)/2 & \text{if } v_{j,i,G} < x_{j,min} \\ \left(x_{j,max} + x_{j,i,G}\right)/2 & \text{if } v_{j,i,G} > x_{j,max} \end{cases} \quad (5)$$

3.1.3 Crossover

DE and SHADE: Binomial crossover operation generates the trial vector $\boldsymbol{u}_{i,G}$ from mutated vector $\boldsymbol{v}_{i,G}$ and current vector $\boldsymbol{x}_{i,G}$. The crossover operation uses compare rule with the threshold CR (6). In DE, this threshold is static, on the other hand, in SHADE its value CR_i is calculated for each individual in generation. CR_i is generated from a normal distribution with a mean parameter value $M_{F,r}$ (selected randomly from the crossover rate historical memory M_{CR}) and standard deviation value of 0.1 (7). If the CR_i value is outside of the interval [0, 1], the closer limit value (0 or 1) is used.

$$u_{j,i,G} = \begin{cases} v_{j,i,G} & \text{if } rand[0,1] \leq CR_i \text{ or } j = j_{rand} \\ x_{j,i,G} & \text{otherwise} \end{cases} \quad (6)$$

$$CR_i = N\left[M_{CR,r}, 0.1\right] \quad (7)$$

The j index is the index of vector feature and j_{rand} is the index of a feature (randomly selected), which has to be taken from the mutated vector. Without the j_{rand} index, the trial vector $\boldsymbol{u}_{i,G}$ could be the same as the current vector $\boldsymbol{x}_{i,G}$ and that would result in unnecessary elitism in the next step of the algorithm.

3.1.4 Elitism

DE and SHADE: Elitism is the algorithm feature which ensures that the next generation G+1 will contain only

equal or better individuals in terms of objective function value (8). If the objective function value of the trial vector $\boldsymbol{u}_{i,G}$ is better than that of the current vector $\boldsymbol{x}_{i,G}$, the trial vector will become the new individual in new generation $\boldsymbol{x}_{i,G+1}$ and the original vector $\boldsymbol{x}_{i,G}$ will be moved to the external archive of inferior solutions A (SHADE only). Otherwise, the original vector remains in the population in next generation and external archive remains unchanged.

$$\boldsymbol{x}_{i,G+1} = \begin{cases} \boldsymbol{u}_{i,G} & \text{if } f(\boldsymbol{u}_{i,G}) < f(\boldsymbol{x}_{i,G}) \\ \boldsymbol{x}_{i,G} & \text{otherwise} \end{cases} \quad (8)$$

3.1.5 Historical Memory Update

SHADE: Values of F_i and CR_i of individuals successful in elitism are stored into two corresponding arrays S_F and S_{CR}. After each generation, those arrays are used to update k-th cell in both historical memories M_F and M_{CR}. The index k is initialized to 1 before the first generation and after each update it is incremented by 1. If it overflows the size of historical memories H, it is set back to 1. When the whole generation fails to improve, S_F and S_{CR} arrays are empty and no update takes place. Also the k index value stays the same. Equations used for historical memory updates are given in (9) and (10).

$$M_{F,k,G+1} = \begin{cases} \text{mean}_{WL}(S_F) & \text{if } S_F \neq \emptyset \\ M_{F,k,G+1} & \text{otherwise} \end{cases} \quad (9)$$

$$M_{CR,k,G+1} = \begin{cases} \text{mean}_{WA}(S_{CR}) & \text{if } S_{CR} \neq \emptyset \\ M_{CR,k,G+1} & \text{otherwise} \end{cases} \quad (10)$$

The weights for both weighted Lehmer mean $\text{mean}_{WL}(S_F)$ and weighted arithmetic mean $\text{mean}_{WA}(S_{CR})$ are evaluated by (11) and used in mean equations given in (12) and (13).

$$w_k = \frac{\text{abs}\left(f(\boldsymbol{u}_{k,G}) - f(\boldsymbol{x}_{k,G})\right)}{\sum_{m=1}^{|S_{CR}|} \text{abs}\left(f(\boldsymbol{u}_{m,G}) - f(\boldsymbol{x}_{m,G})\right)} \quad (11)$$

Since both arrays S_{CR} and S_F are of the same size, either of them can be used for the m index upper boundary of the sum in (11).

$$\text{mean}_{WL}(S_F) = \frac{\sum_{k=1}^{|S_F|} w_k \cdot S_{F,k}^2}{\sum_{k=1}^{|S_F|} w_k \cdot S_{F,k}} \quad (12)$$

$$\text{mean}_{WA}(S_{CR}) = \sum_{k=1}^{|S_{CR}|} w_k \cdot S_{CR,k} \quad (13)$$

3.2 Multi-Chaotic Framework for Parent Selection

Both mutation strategies rand/1 and current-to-$p$best/1 require randomly chosen parents, therefore the mutation can be significantly influenced by the used PRNG for the parent selection. It was experimentally tested that the chaotic map based PRNGs used for parent selection may improve the convergence speed and the ability to reach the global optimum. But the chaotic PRNG which improved the performance of the algorithm on one objective function might not be as suitable as other chaotic PRNG on different objective function. Therefore, the multi-chaotic framework was developed.

Multi-chaotic framework for parent selection presented in this paper was partially inspired by the ranking selection process in Genetic Algorithm (GA) (Holland, 1975). In order to implement framework into the evolutionary algorithm, a chaotic map based PRNG pool ***Cpool*** has to be added to the process. The ***Cpool*** used in this research contains five chaotic PRNGs and each of them has assigned probability value pc_j where j is the index of chaotic PRNG. At the beginning, all pc_j values are initialized to the same pc_{init} value, $pc_{init} = 1/Csize$ where $Csize$ is the size of ***Cpool***. As for this paper, $Csize = 5$ and $pc_{init} = 1/5 = 0.2 = 20\%$. The pc_j value determines the probability of a j-th chaotic PRNG to be used for parent selection.

For each individual vector $\boldsymbol{x}_{i,G}$ in generation G, the chaotic generator $PRNG_k$ is selected from the ***Cpool*** based on its probability pc_k, where k is the index of the selected generator. The selected generator is then used for the random selection of parent vectors. If the trial vector $\boldsymbol{u}_{i,G}$ generated from these parent vectors succeeds in elitism, then the probability pc_k of the selected generator $PRNG_k$ is increased and all other generators probabilities are decreased. The upper boundary for the probability is 60%, $pc_{max} = 0.6$. If the selected chaotic PRNG reaches the maximum probability, then no adjustment takes place. The probability adjustment process is depicted in pseudo-code below – Algorithm 1.

Algorithm 1: Probability adjustment of multi-chaotic system

```
1   Cpool = {Burgers, Delayed Logistic,
    Dissipative, Lozi, Tinkerbell};
2   Csize = 5, pc_max = 0.6;
3   k is the index of the selected
    chaotic system and pc_k is its
    selection probability;
4   if f(u_i,G) < f(x_i,G) and pc_k < pc_max
    then
5   for j = 1 to Csize do
6   if j = k then
7   pc_j = (pc_j + 0.01)/1.01;
8   else
9   pc_j = pc_j /1.01;
10  end
11  end
12  else
13  pc_j = pc_j;
14  end
```

Since the parent selection processes of DE and SHADE differ, the pseudo-code is divided into two algorithms – 2 for DE and 3 for SHADE.

Algorithm 2: Multi-chaotic parent selection in DE

```
1   Cpool = {Burgers, Delayed Logistic,
    Dissipative, Lozi, Tinkerbell};
2   Csize = 5;
3   i is the index of active individual
    x_{i,G};
4   P is the current population of
    individuals;
5   Nsize = |P|;
6   k = 1 the index of selected chaotic
    system;
7   selectedChaos = Burgers;
8   prob = U[0,1];
9   do
10  prob = prob − pc_{chaosIndex};
11  k++;
12  if k > Csize then break;
13  while (prob > 0)
14  k = k - 1;
15  selectedChaos = Cpool[k];
16  do
17  x_{r1,G} = P[selectedChaos.rndInt(1,
    Nsize)];
18  x_{r2,G} = P[selectedChaos.rndInt(1,
    Nsize)];
19  x_{r3,G} = P[selectedChaos.rndInt(1,
    Nsize)];
20  while (x_{i,G} = x_{r1,G} or x_{i,G} = x_{r2,G} or
    x_{i,G} = x_{r3,G} or x_{r1,G} = x_{r2,G} or x_{r1,G} =
    x_{r3,G} or x_{r2,G} = x_{r3,G})
```

Algorithm 3: Multi-chaotic parent selection in SHADE

```
1   Cpool = {Burgers, Delayed
    Logistic, Dissipative, Lozi,
    Tinkerbell};
2   Csize = 5;
3   i is the index of active
    individual x_{i,G};
4   P is the current population of
    individuals, A is the external
    archive, PB is the array of best
    individuals in population, its
    size is given by p in SHADE
    algorithm;
5   Nsize = |P|, NAsize = |P| + |A|,
    PBsize = |PB|;
6   k = 1 the index of selected
    chaotic system;
7   selectedChaos = Burgers;
8   prob = U[0,1];
9   do
10  prob = prob − pc_{chaosIndex};
11  k++;
12  if k > Csize then break;
13  while (prob > 0)
14  k = k - 1;
15  selectedChaos = Cpool[k];
16  do
17  x_{pbest,G} = PB[selectedChaos.rndInt(1,
    PBsize)];
18  x_{r1,G} = P[selectedChaos.rndInt(1,
    Nsize)];
19  x_{r2,G} = (P ∪
    A)[selectedChaos.rndInt(1,
    NAsize)];
20  while (x_{i,G} = x_{pbest,G} or x_{i,G} = x_{r1,G}
    or x_{i,G} = x_{r2,G} or x_{r1,G} = x_{r2,G} or x_{r1,G}
    = x_{pbest,G} or x_{r2,G} = x_{pbest,G})
```

The function *rndInt*(*min*, *max*) generates a random integer from the range [*min*, *max*] according to (1).

DE and SHADE algorithms with multi-chaotic parent selection were labeled MC-DE and MC-SHADE and their pseudo-codes are below in algorithms 4 and 5.

Algorithm 4: MC-DE

```
1   G = 0;
2   Randomly initialized population P =
    (x_{1,G, … , x_{NP,G});
3   Cpool = {Burgers, Delayed Logistic,
    Dissipative, Lozi, Tinkerbell};
4   All values in pc set to pc_{init} =
    0.2;
5   while termination = false do
6   for i = 1 to NP do
7   Multi-chaotic parent selection –
    Algorithm 2;
8   u_{i,G} mutation (2) and crossover (6);
9   if f(u_{i,G}) < f(x_{i,G}) then
10  x_{i,G+1} = u_{i,G};
11  Probability adjustment of multi-
    chaotic system – Algorithm 1;
12  else
13  x_{i,G+1} = x_{i,G};
14  end
15  end
16  G++;
17  end
```

Algorithm 5: MC-SHADE

```
G = 0, k = 1, A = ∅, p_min = 2 / NP;
Randomly initialized population P = (x_1,G, … , x_NP,G);
All values in M_F and M_CR set to 0.5;
Cpool = {Burgers, Delayed Logistic, Dissipative, Lozi, Tinkerbell};
All values in pc set to pc_init = 0.2;
while termination = false do
S_F = ∅, S_CR = ∅;
for i = 1 to NP do
r = U[1, H];
F_i,G = C[M_F,r, 0.1];
CR_i,G = N[M_CR,r, 0.1];
p_i,G = U[p_min, 0.2];
Multi-chaotic parent selection – Algorithm 3;
u_i,G mutation (3) and crossover (6);
if f(u_i,G) < f(x_i,G) then
x_i,G+1 = u_i,G;
x_i,G → A;
F_i,G → S_F, CR_i,G → S_CR;
Probability adjustment of multi-chaotic system – Algorithm 1;
else
x_i,G+1 = x_i,G;
end
end
if |A| > NP then delete random individuals from A;
if S_F ≠ ∅ and S_CR ≠ ∅ then
Update M_F and M_CR;
k++;
if k > H then k = 1;
end
G++;
end
```

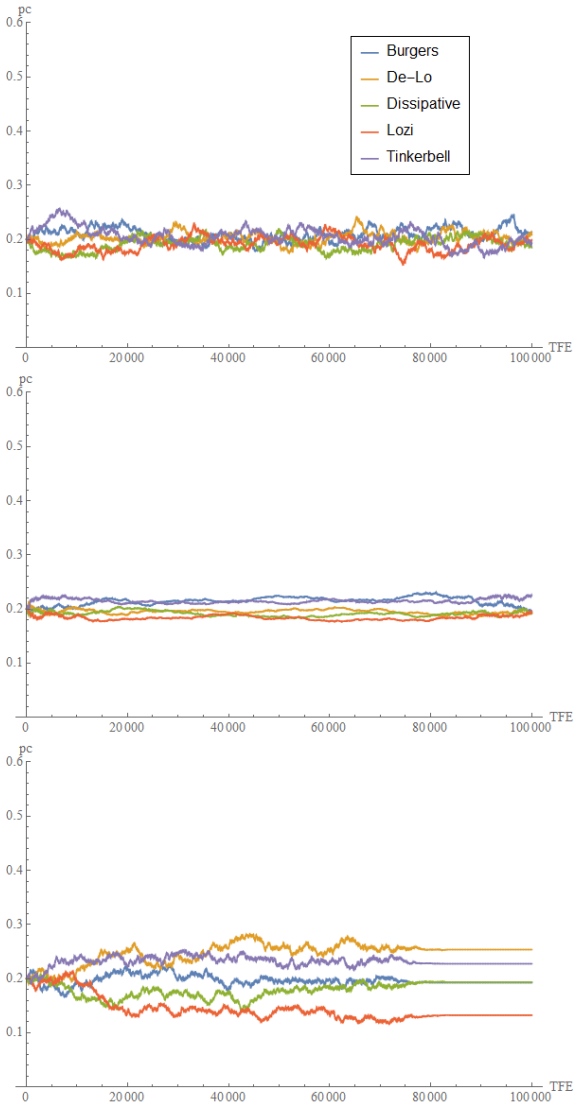


Figure 3. Development of multi-chaotic probabilities over test function evaluations of MC-DE algorithm. From top – Rosenbrock, Rastrigin, Ackley.

4 Experiments and Results

In order to test sensitivity to randomization of non-adaptive and adaptive algorithms, three classic benchmark functions were selected – Rosenbrock (unimodal function with global optima in a narrow, parabolic valley) (14), Rastrigin (complex, multimodal function) (15) and Ackley (multimodal with nearly flat outer region and large hole at the center) (16).

$$f(x) = \sum_{i=1}^{d-1} \left[100\left(x_{i+1} - x_i^2\right)^2 + (x_i - 1)^2\right] \tag{14}$$

$$f(x^*) = 0, x^* = (1, \dots, 1), x_i \in [-2.048, 2.048]$$

$$f(x) = 10d + \sum_{i=1}^{d} \left[x_i^2 - 10\cos(2\pi x_i)\right] \tag{15}$$

$$f(x^*) = 0, x^* = (0, \dots, 0), x_i \in [-5.12, 5.12]$$

$$f(x) = -20e^{-.2\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}} - e^{\frac{1}{d}\sum_{i=1}^{d}\cos(2\pi x_i)} + 20 + e \tag{16}$$

$$f(x^*) = 0, x^* = (0, \dots, 0), x_i \in [-32, 32]$$

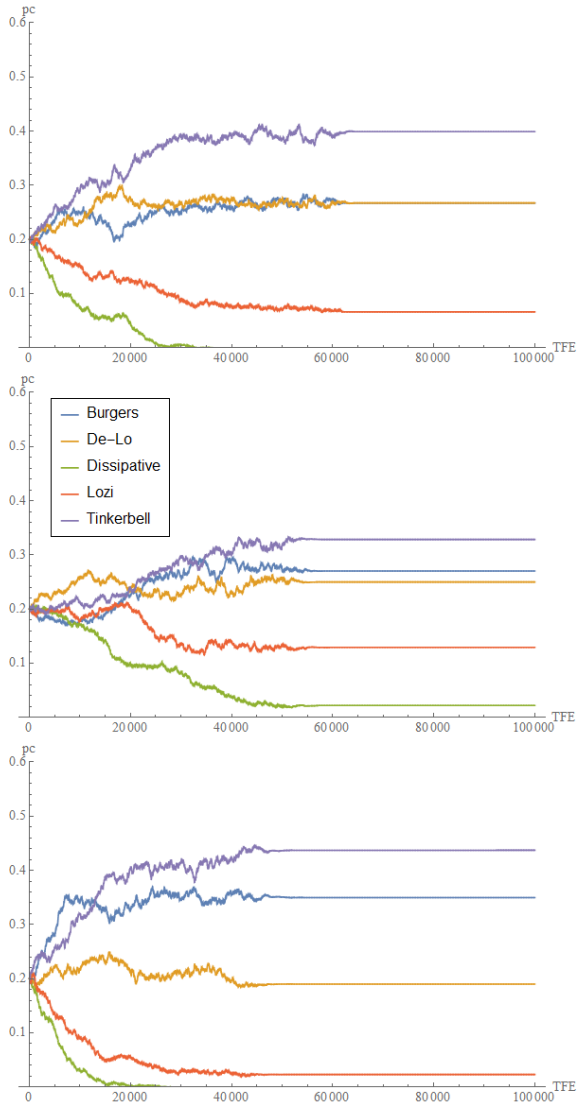


Figure 4. Development of multi-chaotic probabilities over test function evaluations of MC-SHADE algorithm. From top – Rosenbrock, Rastrigin, Ackley.

Both algorithms MC-DE and MC-SHADE were run 51 times on each test function with the maximum number of test function evaluations *maxTFE* set to 100,000. The population size *NP* was set to 100 and 10 dimensional space was selected. The MC-DE control parameters *F* and *CR* were set to 0.5 and 0.8 respectively and MC-SHADEs size of historical memories *H* was set to 10. The multi-chaotic system was initialized with five chaotic maps (Burgers, Delayed Logistic, Dissipative, Lozi and Tinkerbell) with the selection probability pc_{init} = 0.2. The development of the probabilities over the test function evaluations TFE was recorded and all 51 runs were averaged. The average development of probabilities for MC-DE on test functions is shown in Figure 3 and the same is presented for MC-SHADE in Figure 4.

As can be seen in Figure 3, the probabilities of single chaotic maps in non-adaptive algorithm for all three test functions move around the initial values, while as shown in Figure 4, the adaptive algorithm is more sensitive to the randomization system and prefers in each three cases three systems that favor the values close to the left end of specified range for generation. Additionally, in all three cases, the Dissipative chaotic map which was identified to generate values with uniform distribution is strongly suppressed by adaptive algorithm.

5 Conclusions

This paper simulated the effect of adaptivity on randomization on three classic benchmark functions and presented a multi-chaotic framework for parent selection in two DE variants.

In the past, adaptive DE variants outperformed the original DE on numerous benchmarks and real world problems in terms of convergence speed and ability to find the global optimum. Thus, it is important to analyze behavior of such algorithms.

As can be seen in Figure 4, the adaptive algorithm may prefer different randomizations for the selection of parent vectors during the evolutionary process, whereas non-adaptive algorithm seems to be less sensitive and there are mostly none preferred randomizations, which answered the main research question of this paper. The selection of the right randomization or their combination might be beneficial when using adaptive algorithms and the impact has to be studied and analyzed.

The future research will be devoted to thorough analysis of the performance of adaptive algorithms with various randomizations and to development of a robust adaptive randomization system derived from multi-chaotic system presented here.

Acknowledgements

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme Project no. LO1303 (MSMT-7778/2014), further by the European Regional Development Fund under the Project CEBIA-Tech no. CZ.1.05/2.1.00/03.0089 and by Internal Grant Agency of Tomas Bata University under the Projects no. IGA/CebiaTech/2018/003. This work is also based upon support by COST (European Cooperation in Science & Technology) under Action CA15140, Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO), and Action IC1406, High-Performance Modelling and Simulation for Big Data Applications (cHiPSet). The work was further supported by resources of A.I.Lab at the Faculty of Applied Informatics, Tomas Bata University in Zlin (ailab.fai.utb.cz).

References

BV Babu and M Mathew Leenus Jehan. Differential evolution for multi-objective optimization. In *Evolutionary*

Computation, 2003. CEC'03. The 2003 Congress on, volume 4, pages 2696–2703. IEEE, 2003.

Janez Brest, Sao Greiner, Borko Boskovic, Marjan Mernik, and Viljem Zumer. Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE transactions on evolutionary computation*, 10(6):646–657, 2006.

Riccardo Caponetto, Luigi Fortuna, Stefano Fazzino, and Maria Gabriella Xibilia. Chaotic sequences to improve the performance of evolutionary algorithms. *IEEE transactions on evolutionary computation*, 7(3):289–304, 2003.

Nikunj Chauhan, Vadlamani Ravi, and D Karthik Chandra. Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks. *Expert Systems with Applications*, 36(4):7659–7665, 2009.

Leandro dos Santos Coelho, Helon Vicente Hultmann Ayala, and Viviana Cocco Mariani. A self-adaptive chaotic differential evolution algorithm using gamma distribution for unconstrained global optimization. *Applied Mathematics and Computation*, 234:452–459, 2014.

John Henry Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

Hong-Kyu Kim, Jin-Kyo Chong, Kyong-Yop Park, and David A Lowther. Differential evolution strategy for constrained global optimization and application to practical engineering problems. *IEEE Transactions on Magnetics*, 43(4):1565–1568, 2007.

Michal Pluhacek, Roman Senkerik, and Ivan Zelinka. Particle swarm optimization algorithm driven by multichaotic number generator. *Soft Computing*, 18(4):631–639, 2014.

Kenneth Price, Rainer M Storn, and Jouni A Lampinen. *Differential evolution: a practical approach to global optimization*. Springer Science & Business Media, 2006.

Roman Senkerik, Michal Pluhacek, and Zuzana Kominkova Oplatkova. An initial study on the new adaptive approach for multi-chaotic differential evolution. In *Artificial Intelligence Perspectives and Applications,* pages 355–362. Springer, 2015a.

Roman Senkerik, Michal Pluhacek, Zuzana Kominkova Oplatkova, and Donald Davendra. On the parameter settings for the chaotic dynamics embedded differential evolution. In *Evolutionary Computation (CEC), 2015 IEEE Congress on*, pages 1410–1417. IEEE, 2015b.

Julien Clinton Sprott and Julien C Sprott. Chaos and time-series analysis, volume 69. Citeseer, 2003.

Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

Ryoji Tanabe and Alex Fukunaga. Success-history based parameter adaptation for differential evolution. In *Evolutionary Computation (CEC), 2013 IEEE Congress on*, pages 71–78. IEEE, 2013.

Jingqiao Zhang and Arthur C Sanderson. Jade: adaptive differential evolution with optional external archive. *IEEE Transactions on evolutionary computation,* 13(5):945–958, 2009.

Self-adaptive of Differential Evolution using Neural Network with Island Model of Genetic Algorithm

Linh Tao[1] Hieu Pham[2] Hiroshi Hasegawa[3]

[1]D. Functional Control System, Shibaura Institute of Technology, Japan, nb14505@shibaura-it.ac.jp
[2]National Institute of Patent and Technology Exploitation, Vietnam, hieupn@most.gov.vn
[3]D. Functional Control System, Shibaura Institute of Technology, Japan, h-hase@shibaura-it.ac.jp

Abstract

A new evolutionary algorithm called NN-DEGA that using Artificial Neural Network (ANN) for Self-adaptive Differential Evolution (DE) with Island model of Genetic Algorithm (GA) is proposed to solve large scale optimization problems, to reduce calculation cost, and to improve stability of convergence towards the optimal solution. This is an approach that combines the global search ability of DE and the local search ability of Adaptive System with Island model of GA. The proposed algorithm incorporates concept from DE, GA, and Neural Networks (NN) for self-adaptive of control parameters. The NN-DEGA is applied to several benchmark tests with multi-dimensions to evaluate its performance. It is shown to be statistically significantly superior to other Evolutionary Algorithms (EAs), and Memetic Algorithms (MAs).

Keywords: differential evolution, memetic algorithm, migration, neural network, parallel genetic algorithm

1 Introduction

To solve complex numerical optimization problems, researchers have been looking into nature both as model and as metaphor for inspiration. A keen observation of the underlying relation between optimization and biological evolution led to the development of an important paradigm of computational intelligence for performing very complex search and optimization. Evolutionary Computation uses iterative process, such as growth or development in a population that is then selected in a guided random search using parallel processing to achieve the desired end. Nowadays, the field of nature-inspired metaheuristics is mostly continued by the Evolution Algorithms (EAs) (e.g., Genetic Algorithms (GAs), Evolution Strategies (ESs), and Differential Evolution (DE) etc.) as well as the Swarm Intelligence algorithms (e.g., Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC), etc.). Also the field extends in a broader sense to include self-organizing systems, artificial life, memetic and cultural algorithms, harmony search, artificial immune systems, and learnable evolution model. The GAs (Goldberg, 1989; Holland, 1992) have been applied to various complex computational problems, and its validity has been reported by many researchers (Goldberg, 1999; Mahfoud, 1992). However, it requires a huge computational cost to obtain stability in convergence towards an optimal solution. To reduce the cost and to improve the stability, a strategy that combines global and local search methods becomes necessary. As for this strategy, current research has proposed various methods (Nasa, 2016). For instance, Memetic Algorithms (MAs) (Ong, 2004; Smith, 2005) are a class of stochastic global search heuristics in which EAs-based approaches are combined with local search techniques to improve the quality of the solutions created by evolution. MAs have proven very successful across the search ability for multi-modal functions with multi-dimensions (Ong, 2004). These methodologies need to choose suitably a best local search method from various local search methods for combining with a global search method within the optimization process. Furthermore, since genetic operators are employed for a global search method within these algorithms, design variable vectors (DVs) which are renewed via a local search are encoded into its genes many times at its GA process. These certainly have the potential to break its improved chromosomes via gene manipulation by GA operators, even if these approaches choose a proper survival strategy. To solve these problems and maintain the stability of the convergence towards an optimal solution for multi-modal optimization problems with multiple dimensions, Hieu Pham et al. proposed evolutionary strategies of Adaptive Plan system with Genetic Algorithm (APGAs) (Pham, 2012). It is shown to be statistically significantly superior to other EAs and MAs. Unlike most other techniques, GAs maintain a population of tentative solutions that are competitively manipulated by applying some variation operators to find a global optimum. For non-trivial problems, this process might require high computational resources such as large memory and search times. To design efficient GAs, a variety of advances by new operators, hybrid algorithms, termination criteria, and more are continuously being achieved. Parallel GAs (PGAs) (Alba, 1999; Cant, 1998; Tanese, 1989) often leads to superior numerical performance not only to faster algorithms. However, the truly interesting observation is that the use of structured population, either in the form of a set of islands or a diffusion grid, is responsible for such numerical benefits. A PGA has the same as a serial GA, consisting in using rep-

resentation of the problem parameters, robustness, easy customization, and multi-solution capabilities. In addition, a PGA is usually faster, less prone to finding suboptimal solutions only, and able of cooperating with other search techniques in parallel. Differential Evolutionary (DE) was recently introduced and has garnered significant attention in the research literature (Storn, 1997). DE has many advantages including simplicity of implementation, reliable, robust, and in general is considered as an effective global optimization algorithm (Price, 2005). DE operates through similar computational steps as employed by a standard EA. However, unlike traditional EAs, the DE variants perturb the current generation population members with the scaled differences of randomly selected and distinct population members. Therefore, no separate probability distribution has to be used for generating the offspring (Das, 2011). Recently, DE has drawn the attention of many researchers all over the world resulting in a lot of variants of the basic algorithm with improved performance such as Improved Self-adaptive Differential Evolution (ISADE) used in (Bui, 2015) and Advanced DE (ADE) (Mohamed, 2011) etc. (Brest, 2006; Liu, 2005; Mohamed, 2011; Noman, 2008; Omran, 2007; Xu, 2009). Compared with and other techniques (Vesterstroem, 2004), it hardly requires any parameter tuning and is very efficient and reliable. In this paper, we purposed a new evolutionary algorithm called NN-DEGA that using Artificial Neural Network (ANN) for Self-adaptive DE with Island model of GA to solve large scale optimization problems, to reduce a large amount of calculation cost, and to improve the convergence towards the optimal solution.

2 New Evolutionary Computation

2.1 Island model parallel distributed in NN-DEGA

Migration PGA, island model, such as those described in Sect. 1, are reported to have greater information compatibility, a stable design and low computational costs because they deal with GAs in parallel. In NN-DEGA, optimization is conducted by applying GA and DE to each subpopulation. The control variables adjust the vicinity of the output constriction factor F between the subpopulations. The candidate control variables and the new solution come from the other subpopulation at the time of immigration, so a diversity of solutions can be expected because the migration destination is determined at random. A schematic diagram of NN-DEGA with PGA migration is shown in Figure 1.

2.2 Self-adaptive using Neural Network

The self-adaptive constriction factor $F(NN)$ is used for data clustering of the GA control variables using NN, which have been determined uniquely to stabilize their variation. From a viewpoint of excellent parallel processing and to ensure compatibility with multi-point search

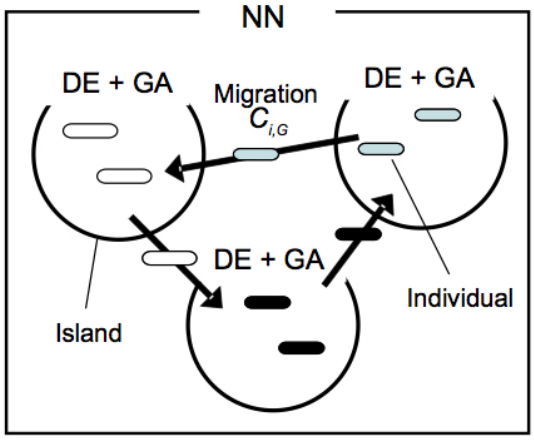


Figure 1. Island Model GA conceptual diagram in NN-DEGA.

methods such as GA and DE are also used in the present method. NN is often used in combination with these techniques (Kobayashi, 2007). NN may be used to cluster and classify the data without using a signal if it is necessary to learn using a teacher signal that is also a NN. In the present method, the GA variable data clustering is controlled using unsupervised learning to determine the output scaling factor change. The initial constriction factor F is set at random and we vary its value based on the NN output. The unsupervised learning method is also a multi-layer NN, so we use NN to perform the feed-forward transfer. The number of layers is determined in a number of search points for each subpopulation. In addition, the NN is configured after it has been sorted in descending order of fitness in the subpopulations to the output side from the input side, where the weight of the transfer equation is as shown in (1). Therefore, many subpopulations have highly adaptive search points with strong effects on other subpopulations. The formulation of the control variable, the transfer equation for each node in the NN and the schematic diagram of the overall NN are as follows.

$$w_{j^n i^n} = y_{nm} / y_{(n-1)m} \tag{1}$$

$$node_t = \sum_{i=1}^{I} SP \cdot w_{i^n j^n} out_i^{n-1} / I \tag{2}$$

$$SP = 2 \cdot C_{i,G} - 1 \tag{3}$$

$$C = [c_{i,j}, \ldots, c_{i,p}]; (0.0 \leq c_{i,j} \leq 1.0) \tag{4}$$

$$F_{i,G+1} = F_{i,G} - \nabla F_i \tag{5}$$

The GA handles control variables (CVs) and C_t is allocated to each search point, which is encoded as a 10-bit string. The order of each search point is allocated to each node of a multi-layer NN, as shown in Figure 2, on the input side and the output side. The weight of the NN, $w_{j^n i^n}$, which is determined from the adaption ratio of the search

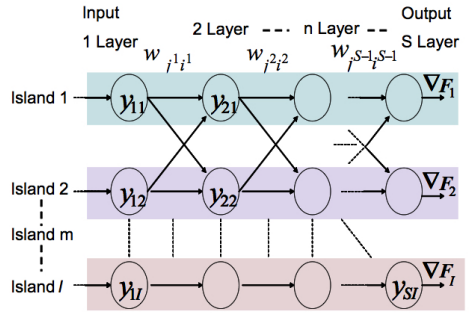


Figure 2. NN-DEGA neural network.

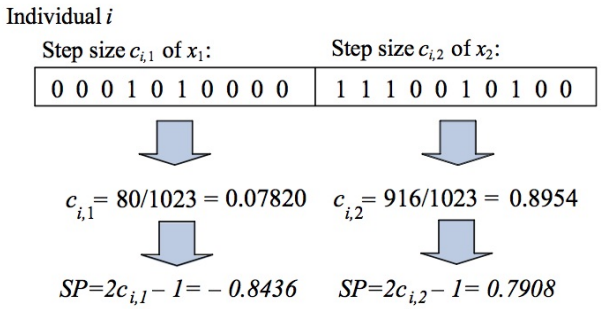


Figure 3. Step size that defined by CVs for controlling a global behavior to prevent it falling into the local optimum.

points, is transmitted between the nodes. C_t is the control variable that determines the step size SP in (3) (as shown in Figure 3) and this element determines the extent of the constraint factor change, ∇F. Therefore, the constriction factor change is an important factor, which determines the width of the overall distribution of the neighborhood of search points. Using the control variable, we can change F adaptively to facilitate more stable solution search and better control of the control variable in the NN. In addition, n is the number of NN hierarchical levels, m is the number of subpopulations, j, i is the number of neurons in NN, t is the number of individuals, S is the number of searches per island and I is the maximum number of islands.

2.3 Reconstruction of differential vector

Each target vector aims at the global optimal solution by updating differential vector based on its best solution has been achieved so far $pbest_{ij}$ and the best solution of all individuals in the population $gbest_j$ (where $j = [1,2,\ldots,D]$, D is the dimension of the solution vector), as following equation:

$$V_{ij,G+1} = gbest_{j,G} + F \cdot \left(pbest_{ij,G} - X_{ij,G}\right) \qquad (6)$$

We carried out the reconstruction of the control variable like considered control variables APGAs (Pham, 2012),

Algorithm 1 The NN-DEGA Pseudocode

```
Initialize population with CVs;
Generate initial DVs;
Evaluate individuals with initial DVs;
while (Termination Condition) do
    Adaptive control of scaling factor F = F(NN) using Neural network;
    Generate DVs via AP with new DE scheme:
    Generate a mutant vector: V_{ij,G+1} = gbest_{j,G} + F(NN) · (pbest_{ij,G} − X_{ij,G});
    Generate a trial vector U_{ij,G+1} through binomial crossover:
    U_{ij,G+1} = { V_{ij,G+1}, (rand_j ≤ CR) or (j = j_rand)
                 { X_{ij,G+1}, (rand_j ≥ CR) and (j ≠ j_rand)   ;
    Evaluate the trial vector U_{i,G};
    if f(U_{i,G}) ≤ f(X_{i,G}) then X_{i,G+1} = U_{i,G} else X_{i,G+1} = X_{i,G};
    end if
    Evaluate individuals with DVs;
    Select parents;
    Recombine to produce offspring for CVs;
    Mutate offspring for CVs;
    if (Restructuring Condition) then
        Restructure chromosome of offspring for CVs;
    end if
end while
```

not only control variable meet the conditions listed below, but also reconstruction of the DE differential vector by keep performing keep the global search of the search point, the appropriate solution search is always performed.

- The same value adaptation accounted for more than 80% for the entire
- The same bit-string chromosome occupies more than 80% for the entire
- The same value of scaling factor accounted for 50% of the total.

2.4 Elite strategy

In this paper, using the diploid genetics is not proper to perform the search using the NN solution (Kouchi, 1992). Generally, GA, information has only a single gene for one individual. However, the structure has a double recessive genetic information that does not appear in the dominant phenotype. Here, in NN, genetic information is treated as a control variable. Information dominance for the NN is elite solution closed to the control variable, as shown in the following equation. With the aim of having a strong influence in the form of dominant inheritance, enhancing the effectiveness of the control variable, advantageously advancing the solution search, elite solution against other sub-populations as the island model of GA.

$$\begin{array}{ll} if\ |eSP - SP_1| - |eSP - SP_2| < 0 & SP = SP_1 \\ if\ |eSP - SP_1| - |eSP - SP_2| > 0 & SP = SP_2 \end{array} \qquad (7)$$

3 Numerical Experiments

In this section, the numerical experiments were performed to compare among strategies. Next, the new algorithm were compared with other techniques for the robustness

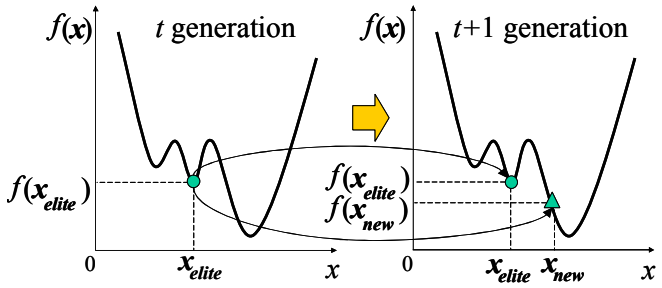


Figure 4. Elite strategy, where the best individual survives in the next generation, is adopted during each generation process.

Table 1. Parameter Settings for Benchmark Tests.

Operator	Control Parameter	Set value
DE	Scaling factor	$F(NN) \in [-1.0, 1.0]$
	Crossover probability	$CR = 0.5$
GA	Selection ratio	1.0
	Crossover ratio	0.8
	Mutation ratio	0.1

The population size: 100

of the optimization approach. These experiments involved 25 independent trials for each function. The parameter settings used in solving the benchmark functions are given in Table 1. The initial seed number are randomly varied during every trial. The scaling factor F takes value in $[-1.0, 1.0]$, and the crossover probability CR is set by 0.5 for the performance of DE. The GA parameters, selection ratio, crossover ratio and mutation ratio are 1.0, 0.8 and 0.01, respectively. The population size has 100 individuals.

3.1 Benchmark Functions

For the NN-DEGA, we estimated the stability of the convergence to the optimal solution by using five benchmarks with 30, and 100 dimensions - Ridge (f_3), Rosenbrock (f_5), Rastrigin (f_9), Ackley (f_{10}), and Griewank (f_{11}). Table 2 lists their characteristics, including the terms epistasis, multi-peaks, and steepness. D denotes the dimensionality of the test problem, design range variables and the global optimum value are summarized. A more detailed description of each function is given in Ref. (Yao, 1999). All functions are minimized to zero, when optimal DVs $X = 0$ are obtained. Note that, it is difficult to search for optimal solutions by applying one optimization strategy only, because each function has specific complex characteristics. The search process is terminated when the search point attains an optimal solution or a current generation process reaches the termination.

$$f_3 = \sum_{i=1}^{D} \left(\sum_{j=1}^{i} x_j \right)^2 \tag{8}$$

Table 2. Characteristics of benchmark tests.

Func	Epis	M-peaks	Steepness	Design range	Optimum
f_3	Yes	No	Average	$[-100, 100]^D$	$f(0) = 0$
f_5	Yes	No	Big	$[-30, 30]^D$	$f(0) = 0$
f_9	No	Yes	Average	$[-5.12, 5.12]^D$	$f(0) = 0$
f_{10}	No	Yes	Average	$[-32, 32]^D$	$f(0) = 0$
f_{11}	Yes	Yes	Small	$[-600, 600]^D$	$f(0) = 0$

D denotes the dimensionality of the test problem.
"Epis" stands for Epistatic, "M - peak" stand for Multi - peaks.

$$f_5 = \sum_{i=1}^{D} [100(x_{i+1} + 1 - (x_i + 1)^2)^2 + x_i^2] \tag{9}$$

$$f_9 = 10D + \sum_{i=1}^{D} [x_i^2 - 10cos(2\pi x_i)] \tag{10}$$

$$f_{10} = -20\exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2}\right) - \exp\left(\frac{1}{D}\sum_{i=1}^{D}\cos(2\pi x_i)\right) + 20 + e \tag{11}$$

$$f_{11} = 1 + \sum_{i=1}^{D} \frac{x_i^2}{4000} - \prod_{i=1}^{D} cos\left(\frac{x_i}{\sqrt{i}}\right) \tag{12}$$

3.2 Experiment Results

The experiment results, average generations required to reach the global optimum of all benchmark functions with 30 dimensions in term of 150,000 FES by the NN-DEGA are given in Tables 3 - 6. "Mean best" indicates average of optimum values obtained and "Std Dev" stands for standard deviation. The solution of all benchmark functions reach their global optimum solutions, and the success rate of optimal solution is 100%. The effect of island number and immigration rate on the performance of algorithm is reported. From this results via optimization experiments, it can be concluded that either the island number or immigration rate increase, the performance of the NN-DEGA algorithm significantly improves. Additionally, the results show that the NN-DEGA algorithm is effective in all benchmarks for various island number and immigration rate, which suggests that the NN-DEGA is more stable and robust on island model using neural network. We employed the best value of island number and immigration rate for the NN-DEGA are 10, 0.2 respectively. In addition, the experiment results with 100 dimension in term of fixed total evaluation times 500,000 FES are given in Table 7. When the success rate of optimal solution is not 100%, "-" is described. We confirmed that the NN-DEGA could solve multi-modal functions with high probability. As a result, its validity confirms that this strategy can dramatically reduce the computation cost and improve the stability of the convergence to the optimal solution more significantly.

Table 3. Experiment results, average generations required to reach the global optimum over 25 runs ($D = 30$, population size 100, island number 5 and imigration rate 0.05).

Function	Gen. No	NFE	Mean best	Std Dev
f_3	571	57,100	0.000E+000	0.000E+000
f_5	147	14,700	0.000E+000	0.000E+000
f_9	61	6,100	0.000E+000	0.000E+000
f_{10}	101	10,100	4.441E-016	0.000E+000
f_{11}	77	7,700	0.000E+000	0.000E+000

Table 4. Experiment results, average generations required to reach the global optimum over 25 runs ($D = 30$, population size 100, island number 5 and imigration rate 0.2).

Function	Gen. No	NFE	Mean best	Std Dev
f_3	380	38,000	0.000E+000	0.000E+000
f_5	95	9,500	0.000E+000	0.000E+000
f_9	45	4,500	0.000E+000	0.000E+000
f_{10}	71	7,100	4.441E-016	0.000E+000
f_{11}	58	5,800	0.000E+000	0.000E+000

Table 5. Experiment results, average generations required to reach the global optimum over 25 runs ($D = 30$, population size 100, island number 10 and imigration rate 0.05).

Function	Gen. No	NFE	Mean best	Std Dev
f_3	585	58,500	0.000E+000	0.000E+000
f_5	129	12,900	0.000E+000	0.000E+000
f_9	61	6,100	0.000E+000	0.000E+000
f_{10}	99	9,900	4.441E-016	0.000E+000
f_{11}	73	7,300	0.000E+000	0.000E+000

Table 6. Experiment results, average generations required to reach the global optimum over 25 runs ($D = 30$, population size 100, island number 10 and imigration rate 0.2).

Function	Gen. No	NFE	Mean best	Std Dev
f_3	394	39,400	0.000E+000	0.000E+000
f_5	83	8,300	0.000E+000	0.000E+000
f_9	43	4,300	0.000E+000	0.000E+000
f_{10}	66	6,600	4.441E-016	0.000E+000
f_{11}	56	5,600	0.000E+000	0.000E+000

Table 7. Experiment results, average generations required to reach the global optimum over 25 runs in term of 500,000 FES ($D = 100$, population size 100, island number 10 and imigration rate 0.2).

Function	Gen. No	NFE	Mean best	Std Dev
f_3	-	-	-	-
f_5	226	22,600	0.000E+000	0.000E+000
f_9	110	11,000	0.000E+000	0.000E+000
f_{10}	232	23,200	4.441E-016	0.000E+000
f_{11}	154	15,400	0.000E+000	0.000E+000

Table 8. Comparison of DE, jDE, ADE and NN-DEGA algorithm in term of 150,000 FES; Gen. No 1500 ($D = 30$, population size=100).

Func.	Gen. No	DE	jDE	ADE	NN-DEGA
		Mean best (Std Dev)	Mean best (Std Dev)	Mean best (Std Dev)	Mean best (Std Dev)
f_3	1500	1.630860 (0.886153)	0.090075 (0.080178)	-	0.000E+000 (0.000E+000)
f_5	1500	7.8E-09 (5.8E-09)	3.1E-15 (8.3E-15)	3.75E-05 (1) (8.90E-05)	0.000E+000 (0.000E+000)
f_9	1500	173.405 (13.841)	1.5E-15 (4.8E-15)	0.0E+00 (2) (0.0E+00)	0.000E+000 (0.000E+000)
f_{10}	1500	9.7E-08 (4.2E-08)	7.7E-15 (1.4E-15)	6.93E-11 (3.10E-11)	4.441E-016 (0.000E+000)
f_{11}	1500	2.9E-13 (4.2E-13)	0 0	0.0E+00 (3) (0.0E+00)	0.000E+000 (0.000E+000)

(1) Gen. No 3000; (2) Gen. No 5000; (3) Gen. No 2000

3.3 Comparison for Robustness

To evaluate the performance of the NN-DEGA algorithm, we compared to other EAs such as GA, PSO, PS-EA in (Srinivasan, 2010), ABC (Karaboga, 2006), DE (Storn, 1997), jDE (Brest, 2006), and ADE (Mohamed, 2011). Maximum number of generation and the population size, i.e. 100, as in the study presented in (Vesterstroem, 2004; Srinivasan, 2010). The mean and the standard deviations of the function values obtained by these methods are given in Tables 9 and 8. By means of the comparison with other methodologies, the NN-DEGA could certainly achieve optimal solution with low calculation cost. Additionally, the results show that the proposed NN-DEGA algorithm outperformed other techniques in all function. The convergence of the optimal solution could be improved more significantly in the NN-DEGA than that in other methods for the same calculation cost. Therefore, it is desirable to introduce this strategy for global optimization.

4 Conclusions

In this paper, overcome the computational complexity, a new evolutionary strategy that using Artificial Neural Network for Self-adaptive Differential Evolution with Island model of Genetic Algorithm called NN-DEGA is proposed to solve large scale optimization problems, to reduce a large amount of calculation cost, and to improve the convergence to the optimal solution. Then, we verified the effectiveness of the NN-DEGA algorithm by the numerical experiments performed five benchmark tests. Moreover, the NN-DEGA was compared to other EAs, it shown to be statistically significantly superior to other EAs. We confirmed that the NN-DEGA reduces the calculation cost and dramatically improves the convergence towards the optimal solution. Moreover, it could solve large scale optimization problems with high probability. About a solution of the problem of cost reduction, minimum time and maximum reliability, it is a future work. Finally, this study plans to do a comparison with the sensitivity plan of the AP by applying other methods on constrained real-parameters and dynamic optimization problems, and further real-life applications.

Table 9. Comparison of GA, PSO, PS-EA, ABC and NN-DEGA algorithm in term of 100,000 FES; Gen. No 1000 ($D = 30$, population size=100).

Func.	Gen. No	GA	PSO	PS-EA	ABC	NN-DEGA
		Mean best (Std Dev)	Mean best (Std Dev)	Mean best (Std Dev)	Mean best (Std Dev)	Mean best (Std Dev)
f_3	1000	-	-	-	-	0.000E+000 (0.000E+000)
f_5	1000	166.283 (59.5102)	402.54 (633.65)	98.407 (35.5791)	0.219626 (0.152742)	0.000E+000 (0.000E+000)
f_9	1000	10.4388 (2.6386)	32.476 (6.9521)	3.0527 (0.9985)	0.033874 (0.181557)	0.000E+000 (0.000E+000)
f_{10}	1000	1.0989 (0.24956)	1.49E-6 (1.86E-6)	0.3771 (0.098762)	3E-12 (5E-12)	4.441E-016 (0.000E+000)
f_{11}	1000	1.2342 (0.11045)	0.011151 (0.014209)	0.8211 (0.1394)	2.87E-09 (8.45E-10)	0.000E+000 (0.000E+000)

References

E. Alba and J.M. Troya. A Survey of Parallel Distributed Genetic Algorithms, *Journal Complexity*, 4(4): 31–52,1999.

J. Brest, S. Greiner, B. Boskovic, M. Mernik, and V. Zumer. Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems, *IEEE Trans. Evol. Comput.*, 10(6): 646–657, 2006.

Ngoc Tam Bui and Hiroshi Hasegawa. Training Artificial Neural Network Using Modification of Differential Evolution Algorithm. *International Journal of Machine Learning and Computing*, 5(1): 1-6, 2015.

E. Cant. A survey of parallel genetic algorithms, *Calculateurs paralleles, reseaux et systems repartis*, vol 10,1998.

S. Das and P.N. Suganthan. Differential evolution - A survey of the State-of-the-Art, *IEEE Transactions on Evolutionary Computation*, 15(1): 4–31, 2011.

D.E. Goldberg. *Genetic Algorithms in Search Optimization and Machine Learning*, Addison - Wesley, 1989.

D. E. Goldberg and S. Voessner. Optimizing global-local search hybrids, *In Proceedings of 1999 Genetic and Evolutionary Computation Conference*, pages 220–228, 1999.

J. Holland. *Adaptation in Natural and Artificial Systems*. The University of Michigan 1975, MIT Press, 1992.

D. Karaboga and B. Basturk. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm, *Journal Global Optimization*, 39: 459–471, 2006.

K. Kobayashi, T. Hiroyasu, and M. Miki. Mechanism of Multi-Objective Genetic Algorithm for Maintaining the Solution Diversity Using Neural Network, *The Science and Engineering Review of Doshisha University*, 48(2): 24–33, 2007.

M. Kouchi, H. Inayoshi, and T. Hoshino. Optimization of Neural-Net Structure by Genetic Algorithm with Diploydi and Geographical Isolation Model, *Japanese Society for Artificial Intelligence*, 7(3): 509–517, 1992.

J. Liu and J. Lampinen. A fuzzy adaptive differential evolution algorithm, *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 9(6): 448–462, 2005.

S. W. Mahfoud and D. E. Goldberg. Parallel recombinative simulated annealing: A genetic algorithm, *Parallel Computing*, 21(1): 1–28, 1995.

A. Wagdy Mohamed, H.Z. Sabry, and A. Farhat. Advanced Differential Evolution algorithm for global numerical optimization, *In IEEE International Conference on Computer Applications and Industrial Electronics (ICCAIE)*, pages 156–161, 2011.

N. Noman and H. Iba. Accelerating differential evolution using an adaptive local Search, *IEEE Transactions on Evolutionary Computation*, 12(1): 107–125, 2008.

Nasa Publications `http://ti.arc.nasa.gov/tech/rse/publications/`

M. Omran, A.P. Engelbrecht, and A. Salman. Empirical analysis of self-adaptive differential evolution, *European Journal of Operations Research*, 183(2): 785–804, 2007.

Y.S. Ong and A.J. Keane. Meta-Lamarckian Learning in Memetic Algorithms, *IEEE Transactions on Evolutionary Computation*, 8(2): 99–110, 2004.

K. Price, R. Storn, and J. Lampinen. *Differential Evolution: A Practical Approach to Global Optimization*, Springer-Verlag, Berlin, 2005.

Hieu Pham, S. Tooyama, and H. Hasegawa. Evolutionary Strategies of Adaptive Plan System with Genetic Algorithm, *JSME Journal of Computational Science and Technology*, 6(3): 129–146, 2012.

D.E. Rumelhart and J.L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, 1986.

D. Srinivasan and T.H. Seow. Evolutionary Computation, *IEEE Congress on Modelling and Simulation*, 2010.

J.E. Smith, W.E. Hart, and N. Krasnogor. *Recent Advances in Memetic Algorithms*, Springer, 2005.

R. Storn and K. Price. Differential Evolution - a simple and efficient Heuristic for global optimization over continuous spaces, *Journal Global Optimization*, 11(4): 341–357, 1997.

R. Tanese. Distributed genetic algorithms, *In Proc. of 3rd Int. Conf. on Genetic Algorithms*, pages 434–439, 1989.

J. Vesterstroem and R. Thomsen. A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems, *In Proc. IEEE Congr. Evolutionary Computation*, pages 1980-1987, 2004.

Y. Xu, L. Wang, and L. Li. An effective hybrid algorithm based on simplex search and differential evolution for global optimization, *In Proc. ICIC*, pages 341–350, 2009.

X. Yao, Y. Liu, and G. Lin. Evolutionary programming made faster, *IEEE Trans. Evol. Comput.*, 3(2): 82–102, 1999.

X. Yao. Evolving artificial neural networks, *In Proceedings of the IEEE*, 87: 1423-1447,1999.

Developing New Solutions for a Reconfigurable Microstrip Patch Antenna by Inverse Artificial Neural Networks

Ashrf Aoad [1] Murat Simsek[2]
[1] Department of Electrical and Electronics Engineering, Istanbul Sabahattin Zaim University, Turkey, ashawad@hotmail.com
[2]Department of Astronautics Engineering, Istanbul Technical University, Turkey, simsekmu@itu.edu.tr

Abstract

This paper presents the use of inverse artificial neural networks (ANNs) to develop and optimize a reconfigurable 5-fingers shaped microstrip patch antenna. New solutions are produced by using three accurate prior knowledge inverse ANNs with sufficient amount of training data where the frequency information is incorporated into the structure of ANNs. The proposed antenna can operate with four modes, which are controlled by two PIN diode switches with ON/OFF states, and it resonates at multiple frequencies between 2-7 GHz. The complexity of the input/output relationship is reduced by using prior knowledge. Three independent methods of incorporating knowledge in the second step of the training process with a multilayer perceptron (MLP) in the first step are demonstrated and their results are compared to EM simulation.

Keywords: artificial neural networks, reconfigurable microstrip antenna, prior knowledge input

1 Introduction

With the rapid development of wireless communication applications, especially in satellites, MIMO systems, radar and portable computers (Costantine et al., 2015). The choice of reconfigurable antennas comes in a large variety of different shapes and forms of the structure. Through change the structure of reconfigurable antennas different characteristics (desired operation) can be obtained (Jiajie and Anguo, 2018; Allayioti and Kelly, 2017). In addition, they have the nature and the capabilities of the reconfiguration mechanism (Aoad et al., 2014). In this application, only ON-ON state of the PIN diode switches is studied (Costantine et al., 2015; Aoad et al., 2014). In a previous study, the structure and results were different (Aoad et al., 2015).

ANN models have been selected for developing new solutions of the proposed antenna. They have the opportunity for modelling and optimizing non-linear relationships between multiple outputs and inputs (Huff and Bernhard, 2008). Therefore, ANNs can be used in the design, development and optimization of antennas (Aoad et al., 2014), integrated circuit-antenna modules and microwave circuits (Wang and Zhang, 1997). The accuracy depends on sufficient training data presented during the training process. In this application, training data is generated by CST-EM simulator.

In this study, two-steps of EM-ANN model is processed using MLP in the first step to model the geometrical dimensions (response) of the proposed antenna, followed by adding extra knowledge into neural networks (NNs) in the second step to correct the response that achieved from first step. Finally, the obtained from second step will be redesigned by EM-simulator to be the developed new solution.

2 Reconfigurable Antenna Design

The studied antenna is a reconfigurable 5-fingers shaped microstrip patch antenna (R5SMPA). This R5SMPA consists of three layers and feeding system at the center of the middle patch. The radiating conductors (first layer) consist of three strips with different dimensions. The parameter of L_1(1.35 cm) and L_2(0.75 cm) are mirrored to the other side where L_3 (1.05 cm) is the middle strip. All strips linked by W_1(3.3 cm) where W_2 (0.3 cm) is the width of all strips. They are positioned on FR-4 dielectric board (second layer) with a thickness of 0.2 cm and ground plane (third layer) is printed on the back side of the dielectric (substrate). W_3 (0.15 cm) is the unfilled space includes two PIN diodes (D_1 and D_2) (ITTC, 2016). They are positioned to distribute the current paths on the microstrips depending on its bias state as shown in Figure 1. To realize the ON-ON state two resistors are used of the PIN diodes (ITTC, 2016). Each resistor has a resistance value of 5 Ohms.

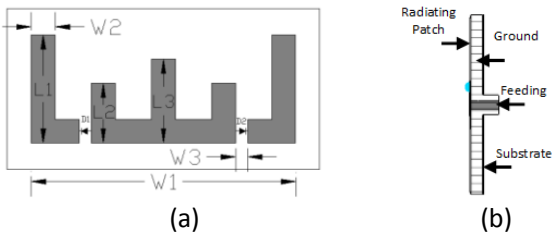


Figure 1. Reconfigurable antenna (a) Top view and (b) side view.

3 Inverse Artificial Neural Networks

For testing the effectiveness of used methods, which were used in (Aoad et al., 2015). Same methods have been applied to the proposed antenna that has a different structure as shown in Figure 1. The proposed inverse ANNs consist of two-steps which is called knowledge based response correction (KBRC), followed by EM simulator to redesign the results of KBRC for obtaining new solutions of the R5SMPA. Input for KBRC model is only frequency sample points, while outputs are considered the new geometrical dimensions of the R5SMPA. It is important to notice that MLP response which is obtained from the first step of KBRC model is not corrected yet. However, SD, PKI-D and PKI models in the second step correct that response.

3.1 Multilayer Perceptron (MLP)

MLP (no extra knowledge) consists of three perceptron layers lined as an input layer, one or more hidden layers and finally an output layer (Zhang and Gupta, 2000) and is in the first step of KBRC model, corresponding to model Y and X variables respectively. The function of the input and the output vectors can be presented as $X = f(Y)$. In this study, the input parameter is $Y_f = [f]^T$ (Y_f presents 200 samples of S-parameters) and the predicted output is $X_c = [L_1, L_2, L_3\ ]^T$.

3.2 Source Difference Method (SD)

The idea of SD (Simsek et al., 2010; Zhang and Gupta, 2000) is in combining two training data sets to be the target of the network. These data sets are the EM simulation outputs of $X_f = [L_1, L_2, L_3\ ]^T$ which represents the fine data and the output response of MLP (X_c) obtained from the first step. Thus, the input parameter of the SD is only the frequency samples $Y_f = [f]^T$, the predicted output $X_{SD} = X_C + X_{MLP}$, while the target is $\Delta X_{SD} = X_{RL} - X_C$. SD is positioned in the second step of KBRC as shown in Figure 2. The function of the input and the output of the redesign case of EM-simulation is presented as

$$Y_{f-SD} = f_{EM}(X_{SD}) \qquad (1)$$

where Y_{f-SD} is the result obtained by the redesign of the predicted output (X_{SD}) of the second step. e_{SD} is the error measure computes the absolute difference between Y_{f-SD} and Y_f which can be calculated by

$$e_{SD} = |Y_{f-SD} - Y_f| \qquad (2)$$

Equations (1) and (2) are same as a general principle in the next methods.

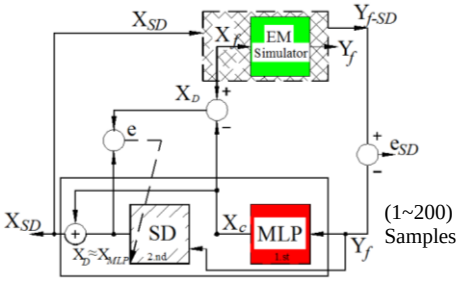


Figure 2. Two steps of KBRC model when SD is in 2nd step of processing.

3.3 Prior Knowledge Input Method (PKI)

In this method (Zhang and Gupta, 2000), the output response of MLP (X_C) is used as input to PKI, in addition to the original input of Y_f. The target output is the fine output (X_f). Therefore, the input/output mapping (in the first step) is between the output response of MLP (X_c) and (Y_f). Thus, the input parameter for PKI is $Y_{PKI} = [\ Y_f, X_C\ ]^T$. PKI is positioned in the second step of KBRC as shown in Figure 3.

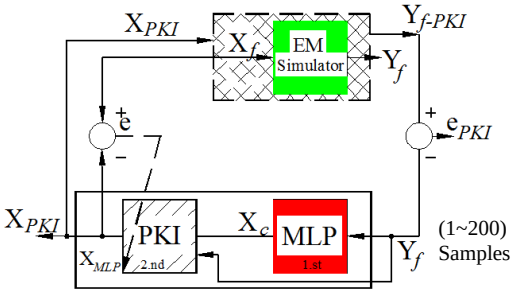


Figure 3. Two steps of KBRC model when PKI is in 2nd step of processing.

3.4 Prior Knowledge Input with Different

PKI-D (Aoad et al., 2014; Zhang & Gupta, 2000) is developed to combine advantages of two knowledge based methods (PKI and SD) described previously. The prior knowledge obtained from the output response of MLP (X_C) is used with the input of the fine model (Y_f) to be the input of PKI-D (Y_{PKI-D}). Therefore, the input parameter is $Y_{PKI-D} = [Y_f, X_C\ ]^T$, when the target output is $\Delta X_{PKI-D} = X_f - X_C$. PKI-D is positioned in the second step of KBRC as shown in Figure 4.

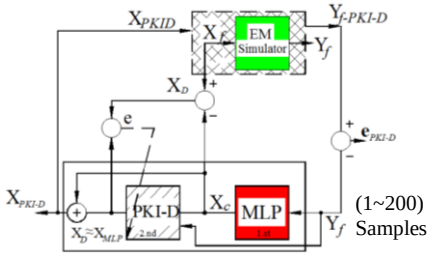


Figure 4. Two steps of KBRC model when PKI-D is in 2nd step of processing.

4 Parameters of ANN Models

Two data sets are initially proposed: 1) Training data and 2) extrapolation testing data sets. The training data generated by EM-simulator was 24800 samples for three geometric antenna parameters ($L_i^k, i = 1,2,3\ and\ k = 5$), k is the number of L_i samples as shown below.

$$N_{tr} = f_s \prod_{i=1}^{3} |L_i^k| \qquad (3)$$

where N_{tr} is the number of training data samples and f_s is the number of frequency samples which is equal to 200. The large amount of training data has been reduced to be 27 samples only (Bataineh and Marler, 2017). The reduction procedure depends on the selection of resonant frequency samples from the training data. The frequency sample points are 200 which are considered the input of the studied models and the outputs are three parameters which are the geometrical dimensions of the R5SMPA. In testing stage, two testing data sets are selected. The testing data sets are selected outside training data which are used to test the accuracy of the models for extrapolation (Simsek et al., 2010). The number of hidden layers is two for all methods. However, the number of neurons is (20-20) for MLP and (30-20) for knowledge based neural networks (KBNNs). Inverse ANN models are trained by using Levenberg-Marguardt algorithm, with tangent-sigmoid transfer functions (TFs) in the hidden layers and a purely linear function in the output layer (Beale et al., 2013). The training parameters of the model are realized by adjusting the learning rate (η) to 0.1 for MLP and 0.05 for KBNNs, the performance goal to 0.000001 for MLP and KBNNs and momentum coefficient (μ) to 0.2 for MLP and 0.1 for others. The regularization coefficient of the network is chosen as 0.2.

5 Results and Discussion

The neural network models are repeatedly trained 50 times and new geometrical dimensions of R5SMPA are developed. The accuracy of the models is presented by the optimum resonant frequency and return loss of the S-parameter curves which are the results of the simulating the new geometrical parameters that obtained by inverse ANN models for extrapolation testing data. Figure 5 shows the result of R5SMPA without training any ANNs as explained in section 2 and shown in Figure 1(a), followed results are by using inverse ANNs.

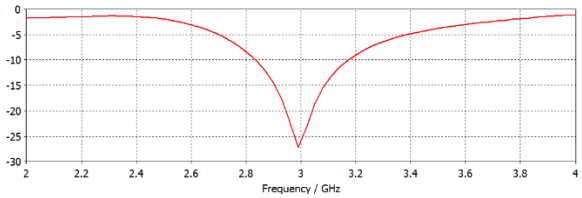


Figure 5. S-parameter for R5SMPA modeled by EM-Simulator, at 3 GHz.

Table 1. A Comparıson Between Results Obtaıned By Inverse ANNs.

Parameters	*Test*	*MLP*	*SD*	*PKI*	*PKI-D*
L_1 (cm)	-	1.4271	1.4738	1.3960	1.4377
L_2 (cm)	-	1.1115	1.0995	1.0818	1.1184
L_3 (cm)	-	0.8535	0.8496	0.8578	0.8340
f_{op} (GHz)	2.44	2.71	2.68	2.74	2.71
RL (dB)	-	-26.89	-32.00	-25.91	-25.73

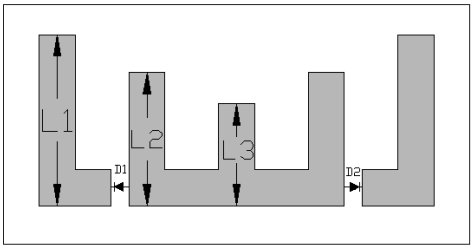


Figure 6. Top view of the developed solution for MLP only at 2.44 GHz.

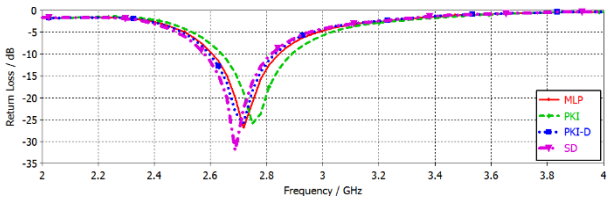


Figure 7. Comparison of S-parameters. The plots show results obtained by ANNs then designed by EM simulator, at 2.44 GHz.

Table 2. A Comparıson Between Results Obtaıned by Inverse ANNs at 3.74 GHz Extrapolation Testing Data.

Parameters	*Test*	*MLP*	*SD*	*PKI*	*PKI-D*
L_1 (cm)	-	0.5521	0.5445	0.5905	0.5466
L_2 (cm)	-	0.7800	0.7923	0.7745	0.7900
L_3 (cm)	-	0.9486	0.9618	0.9647	0.9689
f_{op} (GHz)	3.74	3.77	3.74	3.77	3.77
RL (dB)	-	-21.05	-21.11	-22.47	-22.05

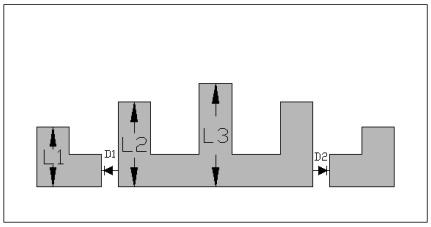


Figure 8. Top view of the developed solution for MLP only, at 3.74 GHz.

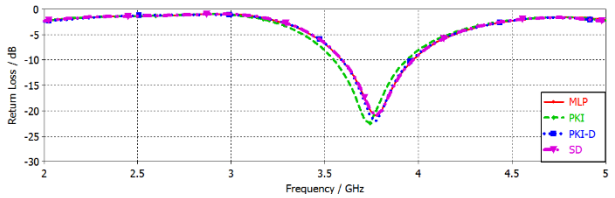


Figure 9. Comparison of S-parameters. The plots show results obtained by ANNs then designed by EM simulator, at 3.74 GHz.

Table 1 and Table 2 show the results of the redesigned physical parameters obtained by inverse ANN models, in addition to optimum frequencies and their return losses. The developed new solutions have new values and shapes for R5SMPA that differ from the original R5SMPA which is shown in Figure 1(a) and the result of Figure 5. Every solution obtained from ANNs operates successfully at a close resonant frequency band as shown in Figure 7 and Figure 9. Moreover, Figure 7, and Figure 9 show that the bandwidths are around 0.32 GHz and 0.41 GHz respectively, at a target of return loss of $S_{11} \leq -10$ dB.

From the results presented above, it is noticed that ANN models are reliable and accurate models**.** The antenna can be reconfigured to obtain new results as well at different extrapolation testing data sets and new switching states such as ON-OFF or OFF-OFF. By comparing results with obtained in (Aoad et al., 2015). The difference was in resonating frequencies, their optimum points, bandwidths, number of input samples and speed of simulations.

6 Conclusions

The proposed inverse ANN methods have been applied for developing new solutions of the reconfigurable antenna. They consist of two steps of processing. MLP is in the first step, SD, PKI-D and PKI are in the second step to correct the response comes from MLP, then the results obtained by the second step redesigned by EM simulator. All methods is applicable to 2 hidden layers. Finally, after applying the proposed inverse ANN methods for developing new solutions of two different shaped reconfigurable antennas. The antenna designers can use same methods to design reconfigurable antennas to reach new solutions and high accuracy.

References

A. Aoad, M. Simsek and Z. Aydin. Design of a Reconfigurable 5-Fingers Shaped Microstrip Patch Antenna by Artificial Neural Networks. *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE),* 4 (10): 61-70, 2014. Available via www.ijarcsse.com

A. Aoad, Z. Aydin and E. Korkmaz. Design of a Tri band 5-Fingers Shaped Microstrip Patch Antenna with an Adjustable Resistor. *Antenna Measurements & Applications (CAMA), IEEE Conference.* Antibes Juan-Les-Pins, 2014.

A. Aoad, M. Simsek and Z. Aydin. Development of Knowledge Based Response Correction for a Reconfigurable N-Shaped Microstrip Antenna Design. *IEEE MTT-S International Conference.* Ottawa, 2015. doi: 10.1109/NEMO.2015.7415078

F. Wang and Q. J. Zhang. Knowledge-Based Neural Models for Microwave Design. *IEEE Transaction on Microwave Theory and Techniques,* 45 (12): 2333-2343, 1997.

C. A. Balanis. *Modern Antenna Handbook*. John Wiley & Sons, Inc. 369-395. 2008.

Information *and Telecommunication Technology Center (ITTC)*. Retrieved 02 25, 2016, from http://www.ittc.ku.edu/~jstiles/622/handouts/old%20handouts/PIN%20Diodes.pdf

J. Costantine, Y. Tawk, S. E. Barbin and C. G. Christodoulou. Reconfigurable Antennas: Design and Applications. *Proceedings of the IEEE,* 103(3): 424-437, 2015.

M. Allayiti and J. R. Kelly. Multiple Parameter Reconfigurable Microstrip Patch Antenna. *IEEE International Symposium,* San Diego. 2017. doi:10.1109/apusncursinrsm.2017.8072613

M. Bataineh and T. Marler. *Neural Networks*. Elsevier. 1-9. 2017.

M. H. Beale, M. T. Hagan and H. B. Demuth. *MatLab Neural Network Toolbox User's Guide,* Natick MA, USA, 2013. Available via http://www.mathworks.com/help/nnet/ref/traingdm.html

M. Simsek, Q. Zhang, H. Kabir and N. Sengor. The recent Developments in Knowledge Based Neural Modelling. *Elsevier-Science Direct,* 1 (1):1321-1330, 2010.

Q. J. Zhang and K. C. Gupta. *Neural networks for RF and microwave design.* Artech House. 2000.

Z. Jiajie, W. Anguo and W. Peng. A survey on Reconfigurable Antennas. *IEEE ICMMT Proceedings.* 2008.

Wind Speed Prediction based on Incremental Extreme Learning Machine

Elizabeta Lazarevska

Faculty of Electrical Engineering and Information Technologies – Skopje, University "Ss. Cyril and Methodius" – Skopje, Macedonia, elizabeta.lazarevska@feit.ukim.edu.mk

Abstract

There are many research papers dealing with wind speed forecasting, since it is necessary in many applications, such as agriculture, modern transportation, and wind energy production. This paper presents an alternative approach to modeling and prediction of wind speed based on extreme learning machine, which is gaining a considerable interest in the scientific and research community at the present. Since the wind speed depends on the atmospheric weather conditions, the wind speed forecast in this research is based on different meteorological data, such as ambient temperature, relative humidity, light intensity, dew point, and atmospheric pressure. The paper presents two neural models for wind speed prediction, based on classic and incremental extreme learning machine, which exhibit the attributes of extreme simplicity, extremely good approximation performance, and extremely fast computation. The performance of the models is validated through their performance indices and compared to other available fuzzy and neural models for wind speed prediction. The paper also addresses the applied modeling techniques and proposes a modification which gives improved results and better approximation performance than the original techniques.

Keywords: wind speed prediction, extreme learning machine, incremental extreme learning machine, random nodes

1 Introduction

The world today widely acknowledges the need for alternative sources of energy to fuel fossils, which are still a primary energy source on Earth (Goldemberg, 2012). There are two main reasons for the intensified efforts and increased research in the field of alternative energy – the exhaustion of fossil fuels, due to the ever-growing energy demands on one hand, and the fact that fossil fuels are not renewable on the other hand, and the pollution caused by utilization of fossil fuels (Casper, 2010). Because of this, mankind is forced to turn to alternative sources of energy such as hydro, solar, wind, and biomass, which are known as renewable energy sources, since they are naturally replenished and are a part of Earth's natural environment (Ehrlich, 2013). Among them the interest, research and investment in solar and wind energy are booming now (Brown et al., 2015). Solar and wind energy are similar in many ways. Both are renewable, both are intermittent, which means that both are not always available like on cloudy (sunless) or calm (windless) days, both are clean energies, which means no environmental pollution. However, the solar energy utilization has excelled at the residential level, which means that the roof-top solar panels can be seen everywhere, while the wind energy technologies have not yet. Nevertheless, wind is much promising, free, renewable, clean, and non-polluting source of energy, which is expected to play much more important role in the future in economics, electricity generation and emission control. As a matter of fact, wind energy is now considered as the fastest growing source and this trend is expected to continue (Walker and Swift, 2015). According to Global Wind Energy Council (GWEC) statistics (http://www.gwec.net), the global cumulative installed wind capacity at the end of 2015 was 432,419 MW, which is 25 times greater than the 17,400 MW installed as of 2000. Industry experts predict that with such pace of growth, one third of the world's energy needs will be met through wind power utilization by the end of 2050.

The bottleneck of wind energy utilization is the time-varying, stochastic, intermittent, and complex nature of wind speed. It is well-known that there is a non-linear cubic relationship between wind speed and the power output of wind turbines (Gasch and Twele, 2012), which means that only a small deviation in wind speed will result in a large deviation in wind power output of the wind turbines. Therefore, it is of utmost importance for wind energy systems to accurately measure and estimate wind speed at a given site (Tamura et al., 2001; Soisuvarn et al., 2013; Yang and McKeogh, 2011). Normally, anemometers are used for measuring wind speed. However, measuring the wind speed is considered the most difficult among various climatological variables. For one, wind farm multiple anemometers must be used since the wind speed varies from one wind turbine to another and for other, the masts for mounting cup anemometers, which are the accepted standard for resource assessment, inevitably

become much taller as wind turbines grow in size, thus making the application of wind anemometers much more expensive. The high cost of wind anemometers discourages their widespread application, which is why engineers replace wind anemometers with digital wind speed estimators for broad applications, such as in wind farms (Kusiak et al., 2010; Mohandes et al., 2011).

Research works on wind speed and/or power prediction have been conducted on different time scale horizons (Soman et al., 2010). Since the time scales used for wind speed predictions are ranging from several minutes to several days, the wind speed forecasting techniques can be grouped into very short, short, medium and long-term methods, as is shown in Table 1. Here in this paper a medium-term wind speed (up to 1 day ahead) prediction has been considered.

Wind speed is a critical feature of wind resources, and since the wind speed values depend on the atmospheric weather conditions, the wind speed forecast in this research is based on different meteorological data, such as ambient temperature, relative humidity, light intensity, dew point, and atmospheric pressure.

2 A Short Overview of the Existing Models for Wind Speed Prediction

Many wind speed estimation methods are presented in the literature as of the present moment (Soman et al., 2010; Stensrud, 2007; Sideratos and Hatziargyrion, 2007; Arjun et al., 2014; Toires et al., 2005; Kavasseri and Seetharaman, 2009; El-Fouly and El-Saadany, 2006; Damousis et al., 2004; Lorenzo et al., 2011; Damousis and Dokopoulos, 2012; Li et al., 2001; Ramasamy et al., 2015; Barbounis et al., 2006; Mohandes et al., 2004; Haque et al., 2012). They can be classified according to different criteria. One such classification according to the adopted prediction period is shown in Table 1. Another classification of wind speed prediction models can be performed based on the applied method, as is shown in Figure 1.

The persistence is the simplest method for wind speed prediction that assumes of a strong correlation between present and future values of wind speed. In other words, it assumes that the future values of wind speed equal the present value. Despite its simplicity, the model is as good as any for short term predictions. Its accuracy decreases rapidly with increasing the prediction time scale.

Table 1. Wind Speed Prediction Time Scales (Soman et al., 2010).

Time Horizon	*Scale*
Very short-term	few seconds up to 30 min ahead
Short-term	30 min up to 6 hours ahead
Medium-term	6 hours up to 1 day ahead
Long-term	1 day up to 5 days ahead

The numerical weather prediction (NWP) models (Stensrud, 2014) use mathematical models of the atmosphere and oceans to predict the weather based on current weather conditions. The complex mathematical calculations involved in modern weather prediction require super powerful computers, and yet, the forecasting ability of NWP does not extend past several days, due to the errors caused by the chaotic nature of the partial differential equations governing the atmosphere.

Statistical prediction methods include regression and time-series models (Sideratos and Hatziargyrion, 2007; Arjun et al., 2014; Toires et al., 2005; Kavasseri and Seetharaman, 2009; El-Fouly and El-Saadany, 2006), which can be linear and non-linear models, structural, grey-box, and black box models. Regression analysis is widely used for prediction and forecasting, and some of the most familiar methods are linear regression, ordinary least squares regression, nonlinear regression etc. The structural models for wind speed and power forecasting include an explicit function of meteorological variable predictions.

Along with the conventional forecasting methods, soft computing methods can also be used for wind speed prediction (Damousis et al., 2004; Lorenzo et al., 2011; Damousis and Dokopoulos, 2012; Li et al., 2001; Ramasamy et al., 2015; Barbounis et al., 2006; Mohandes et al., 2004; Haque et al., 2012). Recent research work has focused on artificial neural networks (ANN) and support vector machines (SVM), which generally produce superior approximation performance compared to other forecasting techniques. The wind speed models based on artificial intelligence techniques,

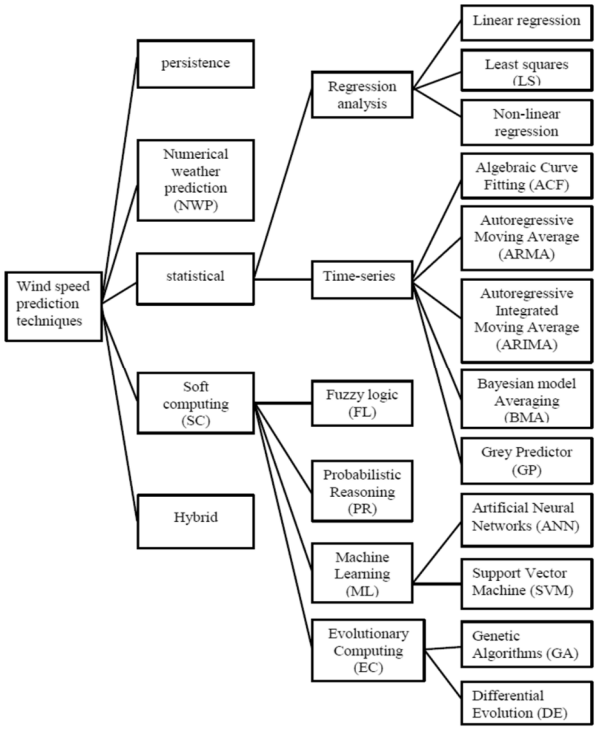


Figure 1. Wind speed prediction techniques.

such as ANN and SVM, belong to the class of black-box models. On the other hand, the expert systems based on fuzzy logic techniques, belong to the class of grey-box models. Hybrid models for short-term wind speed prediction have been presented in (Haque et al., 2012; Shi et al., 2013). They unite different modeling techniques and approaches, such as NN and genetic algorithms (GA), or NN and fuzzy inference systems (FIS).

Much more insight into different techniques for wind speed prediction can be found in (Bhaskar et al., 2010; Lawan et al., 2014; Lei et al., 2009).

3 A Brief Overview of Extreme Learning Machine

The extreme learning machine (ELM) is a new concept in the field of machine learning, which was introduced in (Huang et al., 2004). It was proposed as a new learning algorithm for training feedforward neural networks with single hidden layer (SLFFNN) which is different from the traditional gradient-descent based learning algorithms in the way of adjusting the network parameters. Unlike traditional learning algorithms, that iteratively adjust and tune all the network parameters during the learning process, ELM adjusts only the weights of the connections between the hidden and the output network layer, while the weights of the connections between the input and the hidden layer are assigned randomly at the beginning of the learning process and never updated. As a result, the training of a SLFFNN with ELM is much faster than the training of these networks with traditional gradient-descent based learning methods such as the error back-propagation method (BP).

A classic ELM is a SLFFNN, which has an input layer, one hidden layer and an output layer. For convenience, the weights of the connections between the input neurons and the hidden neurons can be called as input weights, and the weights of the connections between the hidden neurons and the output neurons can be called as output weights. The main features of ELM can be summarized as follows:

- The input parameters of the hidden layer, i.e. the input weights and the biases of the hidden nodes, are randomly assigned according to any continuous probability distribution and fixed during the complete learning process.
- The output parameters of the hidden layer, i.e. the output weights of the hidden nodes, are the only parameters that are learned during the training process.
- The output nodes of ELM do not possess biases.
- ELM is a model that is linear in the parameters.
- SLFFNN with ELM is a universal approximator.
- The hidden nodes activation functions can be almost any non-linear piece-wise continuous functions.
- The hidden nodes may have different output functions.

For a given input-output data set $D = \{(\mathbf{x}_k, \mathbf{y}_k)\}$; $k = 1,2,\cdots,N$ with N distinct data points, where $\mathbf{x}_k = [x_{k1}, x_{k2}, \cdots, x_{kM}]^{\mathrm{T}}$ is the $k-$th input vector of dimension M, and $\mathbf{y}_k = [y_{k1}, y_{k2}, \cdots, y_{kL}]^{\mathrm{T}}$ is the $k-$th output vector of dimension L, a standard SLFFNN with ELM and n hidden neurons can be modeled as:

$$\tilde{y}_k = \sum_{i=1}^{n} \mathbf{v}_i \, g_i(\mathbf{x}_k) = \sum_{i=1}^{n} \mathbf{v}_i \, g(\mathbf{w}_i \mathbf{x}_k + b_i); \quad k = 1,2,\cdots,N \tag{1}$$

The vector $\mathbf{w}_i = [w_{i1} \quad \cdots \quad w_{iM}]^{\mathrm{T}}$ in (1) represents the $i-$th hidden neuron input weights, i.e. the weights defining the connections between the M input neurons and the $i-$th hidden neuron, the vector $\mathbf{v}_i = [\mathbf{v}_{i1} \quad \cdots \quad \mathbf{v}_{iL}]^{\mathrm{T}}$ represents the $i-$th hidden neuron output weights, i.e. the weights defining the connections between the $i-$th hidden neuron and the L output neurons, b_i is the threshold, i.e. the bias of the $i-$th hidden neuron, and $g_i(\mathbf{x}_k)$ is some appropriate activation function of the $i-$th hidden neuron; the term $\mathbf{w}_i \mathbf{x}_k$ in (1) denotes the inner product between the vector of hidden layer input weights $\mathbf{w}_i$ and the input vector $\mathbf{x}_k$.

The Equation (1) can be written in a short matrix form as:

$$\tilde{\mathbf{Y}} = \mathbf{W} \cdot \mathbf{V} \tag{2}$$

where $\tilde{\mathbf{Y}}$ is the model output matrix:

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \vdots \\ \tilde{\mathbf{y}}_N \end{bmatrix} = \begin{bmatrix} \tilde{y}_{11} & \cdots & \tilde{y}_{1L} \\ \vdots & \ddots & \vdots \\ \tilde{y}_{N1} & \cdots & \tilde{y}_{NL} \end{bmatrix} \tag{3}$$

the matrix $\mathbf{W}$ is called a hidden layer output matrix (Huang and Babri, 1998; Huang, 2003):

$$\mathbf{W} = \begin{bmatrix} g(\mathbf{w}_1 \mathbf{x}_1 + b_1) & \cdots & g(\mathbf{w}_n \mathbf{x}_1 + b_n) \\ \vdots & \ddots & \vdots \\ g(\mathbf{w}_1 \mathbf{x}_N + b_1) & \cdots & g(\mathbf{w}_n \mathbf{x}_N + b_n) \end{bmatrix} \tag{4}$$

and $\mathbf{V}$ is the model adjustable parameter matrix:

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_N \end{bmatrix} = \begin{bmatrix} v_{11} & \cdots & v_{1L} \\ \vdots & \ddots & \vdots \\ v_{n1} & \cdots & v_{nL} \end{bmatrix} \tag{5}$$

Because ELM assigns randomly the hidden layer input weights $\mathbf{w}_i$ $(i = 1,2,\cdots,n)$ and the hidden neuron biases b_i $(i = 1,2,\cdots,n)$ and does not adjust them further, the hidden neuron output matrix $\mathbf{W}$ can be calculated only once at the beginning of the ELM learning process and remains fixed throughout the rest of the network training. Reference (Huang et al., 2004) has offered a proof that the input weights and the biases of the hidden neurons in a SLFFNN can be randomly

assigned if the activation function of the hidden layers is infinitely differentiable. Then, the hidden neuron output weights can be determined simply and analytically through the well-known least square methods, since the SLFFNN with fixed hidden neuron input weights and biases becomes a system linear in the parameters and can be solved with adequate techniques for such systems.

The training process of a SLFFNN with fixed hidden layer input weights and fixed hidden neuron biases amounts to finding a least square solution $\mathbf{V}$ of the linear system:

$$\mathbf{Y} = \mathbf{W} \cdot \mathbf{V} \tag{6}$$

where $\mathbf{Y} = [\mathbf{y}_1 \ \cdots \ \mathbf{y}_N]^{\mathrm{T}}$ is the vector of measured output data. When the number of hidden layer neurons n equals the number of input-output data N, the matrix $\mathbf{W}$ is a square matrix and (6) can be solved for $\mathbf{V}$ directly by inversion of $\mathbf{W}$, with a zero error. On the other hand, if the number of hidden neurons n does not equal the number of training data points N, which is always the case in real applications where $n << N$, the matrix $\mathbf{W}$ is not a square matrix and the system (6) is solved for $\mathbf{V}$ with an adequate least square method yielding the smallest norm least square solution (Huang et al., 2004):

$$\mathbf{V} = \mathbf{W}^{+} \cdot \mathbf{Y} \tag{7}$$

In (7) the matrix $\mathbf{W}^{+}$ stands for the Moore-Penrose generalized inverse of matrix $\mathbf{W}$ (Rao and Mitra, 1971; Serre, 2002). The authors in (Huang et al., 2004) have argued that the least square solution (7) is unique, has the smallest norm of weights among all the least square solutions of (6) and secures a minimum training error of ELM.

4 A Brief Overview of Incremental Extreme Learning Machine

The approximation capacity of ELM depends on the number of hidden neurons n. To solve this problem, it has been proposed to gradually increment the number of hidden neurons in the ELM, and this new learning algorithm, presented in Figure 2, has been called as incremental ELM (IELM) (Huang et al., 2006). The difference between the basic ELM and IELM is in the addition of new neurons to the hidden layer of the later. The new neurons can be added one at a time, or in groups, and the process of learning continues until the preset maximum number of hidden neurons is reached, or the preset acceptable model error is achieved. As with the basic ELM, the input parameters of the hidden layer in IELM are randomly generated and need not to be adjusted at all during the learning process (Huang et al., 2006). IELM is very different from other incremental algorithms proposed for SLFFNN with additive hidden nodes (Meir and Maiorov, 2000), since it performs as a universal approximator without tuning the hidden layer output parameters weights. In other words, when a new hidden neuron is added to the hidden layer, the IELM algorithm does not recalculate the hidden layer output parameters of the existing hidden nodes; once they have been determined, they remain fixed throughout the learning process. IELM just calculates the output weight for the newly added hidden neuron according to the following expression:

$$\mathbf{v}_n = \frac{\mathbf{E} \cdot \mathbf{G}^{\mathbf{T}}}{\mathbf{H} \cdot \mathbf{G}^{\mathbf{T}}} = \frac{\sum_{i=1}^{N} e(i) g(i)}{\sum_{i=1}^{N} g^2(i)} \tag{8}$$

The term $g(i)$ in (8) denotes the activation of the added new hidden node for the $i-$th training input sample, while $e(i)$ is the corresponding residual error before the addition of the new hidden node in question; $\mathbf{G} = [g(1) \ \cdots \ g(N)]^{\mathrm{T}}$ is the activation vector for the new added node for all the training samples, and $\mathbf{E} = [e(1) \ \cdots \ e(N)]^{\mathrm{T}}$ is the vector of the residual error before the addition of the new hidden neuron. The new value of the residual error after the addition of a new hidden neuron is calculated according to (Huang et al., 2006):

$$\mathbf{E} = \mathbf{E} - \mathbf{v_n} \cdot \mathbf{G_n} \tag{9}$$

Instead of the IELM algorithm described above, we have tested a much simpler algorithm which is a modification of the basic ELM in the sense that it accepts increasing number of hidden nodes and searches for the smallest preset error defined by ε. The algorithm recalculates the output parameters of all the hidden nodes in the hidden layer after every new addition to the hidden layer and performs until the preset maximum number of hidden neurons or the preset desired model error is reached. The algorithm has been compared to the other algorithms in Table 2 and has shown the best

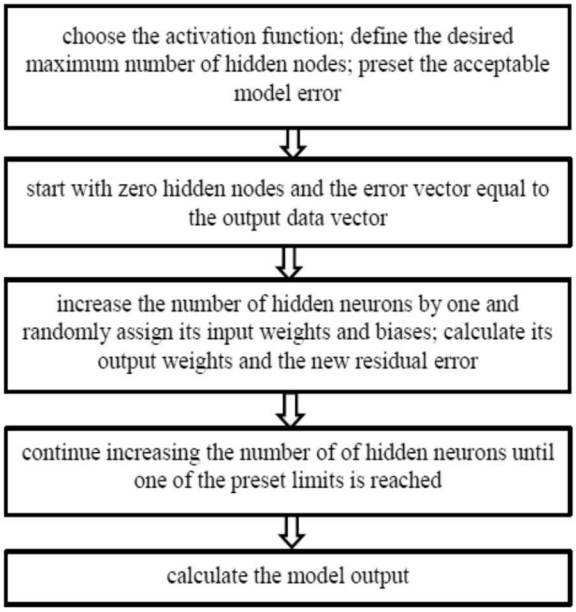


Figure 2. IELM Learning algorithm.

performance in this case. The reason for this lays in the fact that all the output parameters of the hidden layer are recalculated after each new addition of hidden neurons, but this is something that still needs to be examined further.

5 Results and Discussion

All the models for wind speed prediction in this research were built upon the hourly meteorological data for Mauna Loa (MOA), Hawaii, US for year 2015, available at (http://www.nrel.gov/gis/data_wind.html).

The research considering the wind speed forecast was conducted in several directions. First the desired model of the wind speed was obtained by classic ELM with several different activation functions for the hidden neurons. To perform the modeling, the available experimental data were divided into two sets: training data and testing data, and fixed number of hidden nodes were assigned for the ELM algorithm. The input parameters of the hidden neurons $\mathbf{w}_i \ \left(i = 1,2,\cdots,n\right)$ and $b_i \ \left(i = 1,2,\cdots,n\right)$ were randomly assigned according to the uniform probability distribution. The obtained results were compared according the RMSE criterion and the experimental findings are in complete agreement with the claim that ELM produces good generalization performance with almost any non-zero nonlinear activation function (Huang et al., 2004; Huang and Babri, 1998). However, they contradict the claim in (Liu et al., part I, 2015) that "…the performance of ELM depends heavily on the activation function…"

The other group of simulations were performed with the same activation function and a fixed number of hidden neurons, to illustrate the approximation performance of the applied ELM about different random assignments of the hidden layer input parameters. These simulations undoubtedly showed that the approximation performance of classic ELM depends on the randomness mechanism. The random assignment of some ELM parameters causes an uncertainty problem with classic ELM, since there is not a way of knowing whether the obtained model is the best possible solution, i.e. what is the best random assignment of the specific ELM parameters. This dependence was not explicitly and/or clearly mentioned in any of the consulted references at the end of this paper, except in (Liu et al., part I, 2015) to be addressed in (Liu et al., part II, 2015). However, the conducted research in this work showed that the approximation property of ELM suffers from the randomness of the ELM with any type of activation function for the hidden neurons, and not only with Gaussian-type activation functions as concluded in (Liu et al., part II, 2015).

To solve the above problem, in this research the following simple approach is used. After the selection of appropriate activation function and the number of hidden neurons, the desired value of the model error is set as ε. Then the training of the constructed SFFNN with ELM is conducted and the error of the obtained model is compared to the preset desired value ε. If the model error is greater than ε, the training process is repeated. Otherwise, it is considered that the obtained model has the desired accuracy. The model produced in this way, and shown in Figure 3, had much smaller error than the models obtained in the other trials.

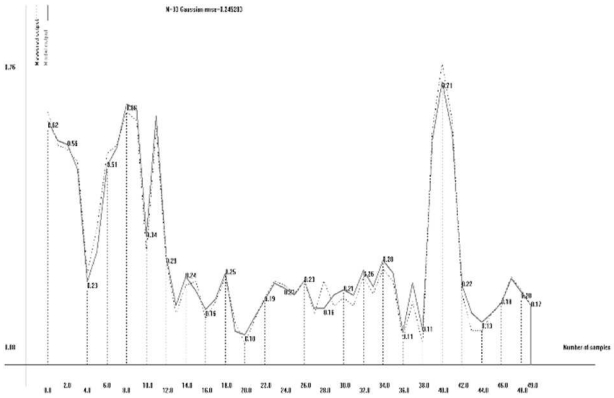


Figure 3. NN model for prediction of wind speed based on our version of ELM; n=33, Gaussian, RMSE=0.245283.

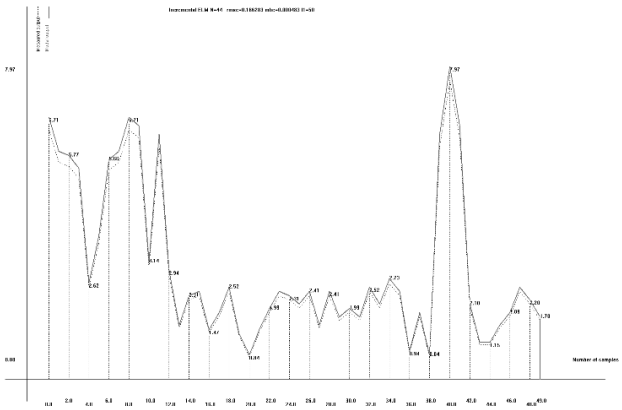


Figure 4. NN model for prediction of wind speed based on IELM; n=44, Gaussian, RMSE=0.166283.

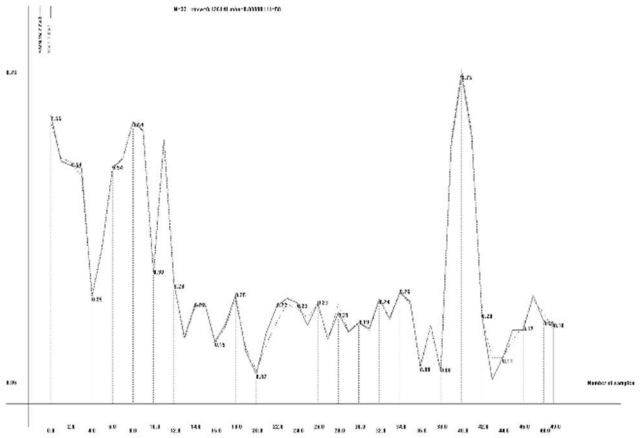


Figure 5. NN model for prediction of wind speed based on our version of IELM; n=33, Gaussian, RMSE=0.126141.

Finally, the wind speed prediction models obtained with our version of ELM, incremental ELM and with our version of incremental ELM were compared to some other models obtained with different modeling techniques, such as the position type and position-gradient type fuzzy models based on Sugeno and Yasukawa identification method (Sugeno and Yasukawa, 1993), the RVM based NN model (Tipping, 2001), and the neuro-fuzzy model based on extended RVM (Kim et al., 2006), and presented in (Lazarevska, 2016). The results are shown in Table 2. All the models are compared according to the RMSE criterion, defined as follows

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2} \qquad (10)$$

where y_i is the measured output, $\tilde{y}_i$ is the model output, and N is the number of experimental data. Figure 3 through Figure 5 shows the output (solid line) of the wind speed prediction models from Table 2 obtained with different ELM algorithms compared to the actual measured output (dashed line). The neural model based on our modification of IELM has the best performance index, since it takes into consideration the randomness factor in ELM.

Table 2. Comparison of Wind Speed Prediction Models obtained by Different Modeling Techniques.

Model	*RMSE*
Position type fuzzy model	0.37940
Position-gradient type fuzzy model	0.32747
RVM based neural model	0.342874
Neuro-fuzzy model based on RVM	0.322158
NN model based on our version of ELM	0.245283
NN model based on IELM	0.166283
NN model based on our version of IELM	0.126141

Acknowledgements

The author gratefully acknowledges that the wind speed prediction models in this research have been based on the wind data available from the National Renewable Energy Laboratory NREL (http://www.nrel.gov/gis/data_wind.html) which is a national laboratory of the US Department of Energy, Office of Energy Efficiency and Renewable Energy, operated by the Alliance for Sustainable Energy (LLC).

References

N. N. Arjun, V. Prema, D. K. Kumar, P. Prashanth, V. S. Preekshit and K. U. Rao. Multivariate regression models for prediction of wind speed. In *2014 International Conference on Data Science and Engineering (ICDSE)*, Kochi 26-28 Aug. 2014, pp. 171-176.

T. G. Barbounis, J. B. Theocharis, M. C. Alexadis and P. S. Dokopoulos. Long term wind speed and powercasting using local recurrent neural network models. *IEEE Transactions on Energy Conversion*, 21(1): 273-284, 2006.

M. Bhaskar, A. Jain and N. V. Srinath. Wind speed forecasting: Present Status. In *2010 International Conference on Power System Technology (POWERCON)*, 24-28 Oct. 2010, Hangzhou, China, 1-6.

L. R. Brown, E. Adams, J. Larsen and J. M. Roney. *The Great Transition: Shifting from Fossil Fuels to Solar and Wind Energy*, 1st edition. W. W. Norton & Company. 2015.

J. K. Casper. *Fossil fuels and pollution: The future of air quality (Global warming (Facts on File))*, 1st edition. Facts on File. 2010.

I. G. Damousis and P. Dokopoulos. A fuzzy expert system for the forecasting of wind speed and power generation in wind farms in *Wind Energy Conversion Systems: Technology and Trends*. S. M. Mayeen ed., London: Springer, 2012, 197-226.

I. G. Damousis, M. C. Alexiadis, J. B. Theocharis and P. S. Dokopoulos. A fuzzy model for wind speed prediction and power generation in wind parks using spatial correlation. *IEEE Transactions on Energy Conversation*, 19(2): 352–361, 2004.

R. Ehrlich. *Renewable energy: A first course*, 1st Edition. CRC Press. 2013.

T. M. H. El-Fouly and E. F. El-Saadany. Grey predictor for hourly wind speed and power forecasting. *IEEE Transactions on Power Systems*, 21(3), 1450-1452, 2006.

R. Gasch and J. Twele. *Wind Power Plants: Fundamentals, Design, Construction, and Operation*, 2nd edition. Springer. 2012.

J. Goldemberg. *Energy: What everyone needs to know®*, 1st edition. Oxford University Press. 2012.

GWEC," *Global wind statistics 2015*," GWEC, 2015. Available at: http://www.gwec.net/wp-content/uploads/vip/GWEC-PRstats-2015_LR_corrected. pdf (accessed 06.01.2016)

A. U. Haque, P. Mandal, J. Meng, M. E. Kaye and L. Chang. A new strategy for wind speed forecasting using hybrid intelligent models. In *2012 25th IEEE Canadian Conference on Electrical and Computer Engineering CCECE*, 2012, 1-4.

G. –B. Huang. Learning capability and storage capacity of two hidden-layer feedforward networks. *IEEE Transactions on Neural Networks*, 14(2): 274-281, 2003.

G. –B. Huang and H. A. Babri. Upper bounds on the number of hidden neurons in feedforward networks with arbitrarily bounded nonlinear activation functions. *IEEE Transactions on Neural Networks*, 9(1): 224-229, 1998.

G. –B. Huang, L. Chen and C. -K. Siew. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions on Neural Networks*, 17(4): 879-892, 2006.

G. –B. Huang, Q. –Y. Zhu and C. –K. Siew. Extreme learning machine: A new learning scheme of feedforward neural networks. In *Proceedings of the IEEE International Joint*

Conference on Neural Networks (IJCNN2004), 25-29 July, 2004, Budapest, Hungary, 985-990.

R. G. Kavasseri and K. Seetharaman. Day-ahead wind speed forecasting using f-ARIMA models. *IEEE Transactions on Renewable Energy*, 34(5): 1388-1393, 2009.

J. Kim, Y. Suga and S. Won. A new approach to fuzzy modeling of nonlinear dynamic systems with noise: relevance vector learning mechanism. *IEEE Trans. on Fuzzy Systems*, 14(2): 222–231, 2006.

A. Kusiak and W. Li. Estimation of wind speed: a data-driven approach. *Journal of Wind Engineering and Industrial Aerodynamics*, 98(10-11): 559-567, 2010.

S. M. Lawan, W. A. W. Z. Abidin, W. Y. Chai, A. Baharun and T. Maasri. Different models of wind speed prediction: A comprehensive review. *International Journal of Scientific and Engineering Research*, 5(1): 1760- 1768, 2014.

E. Lazarevska. A neuro-fuzzy model for wind speed prediction based on statistical learning theory. *Journal of Electrical Engineering and Information Technologies*, 1(1-2): 45-55, 2016.

M. Lei, L. Shiyan, J. Chuanwen, L. Hongling and Z. Yan. A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews*, 13(4): 915-920, 2009.

S. Li, D. C. Wunsch, E. A. O'Hair and M. G. Giesselmann. Using neural networks to estimate wind turbine power generation. *IEEE Trans. Energy Convers*., 16(3): 276–282, 2001.

J. Lorenzo, J. Méndez, M. Castrillón and D. Hernández. Short-term wind power forecast based on cluster analysis and artificial neural networks. In *Proceedings of the 11th International Work-Conference on Artificial Neural Networks, IWANN 2011*, Torremolinos-Málaga, Spain, June 8-10, 2011, Part I, 191-198.

X. Liu, S. Lin, J. Fang and Z. Xu. Is extreme learning machine feasible? A theoretical assessment (Part I). *IEEE Transactions on Neural Networks and Learning Systems*, 26(1): 7-20, 2015.

X. Liu, S. Lin, J. Fang and Z. Xu. Is extreme learning machine feasible? A theoretical assessment (Part II). *IEEE Transactions on Neural Networks and Learning Systems*, 26(1): 21-34, 2015.

R. Meir and V. E. Maiorov. On the optimality of neural network approximation using incremental algorithms. *IEEE Transactions on Neural Networks*, 11(2):323-337, 2000.

M. A. Mohandes, T. O. Halawani, S. Rehman and A. A. Hussain. Support vector machine for wind speed prediction. *Renewable Energy*, 29(6): 939-947, 2004.

M. Mohandes, S. Rehmanand S. M. Rahman. Estimation of wind speed profile using adaptive neuro-fuzzy inference system (ANFIS). *Applied Energy*, 88(11): 4024-4032, 2011.

http://www.nrel.gov/gis/data_wind.html (accessed 06.01.2016)

P. Ramasamy, S. S. Chandel A. K. Yadav. Wind speed prediction in the mountainous region of India using an artificial neural network model. *Renewable Energy*, 80: 338-347, 2015.

C. R. Rao and S. K. Mitra. *Generalized Inverse of Matrices and its Applications*. New York: Wiley. 1971.

D. Serre. *Matrices: Theory and Applications*. New York: Springer-Verlag. 2002.

N. Shi, S. –Q. Zhou, X. –H. Zhu, X. –W. Su and X. –Y. Zhao. Wind speed forecasting based on grey predictor and genetic neural network models. In *2013 International Conference on Measurement, Information and Control (ICMIC)*, 16-18 Aug. 2013, 2:1479 – 1482.

G. Sideratos and N. Hatziargyriou. An advanced statistical method for wind power forecasting. *IEEE Transactions on Power Systems*, 22(1): 258-265, 2007.

S. Soisuvarn, Z. Jelenak, P. S. Cheng, S. O. Alsweiss and Q. Zhu. CMOD5.H. - A high wind geophysical model function for c-band vertically polarized satellite scatterometer measurements. *IEEE Transactions on Geoscience and Remotesensing*, 51(6): 3741-3760, 2013.

S. S. Soman, H. Zareipour, O. Malik and P. Mandal. A review of wind power and wind speed forecasting methods with different time horizons. *North American Power*, 2(5): 8-16, 2010.

D. J. Stensrud. *Parametrization Schemes: Key to understanding numerical weather prediction models*. Cambridge University Press. 2007.

M. Sugeno and T. Yasukawa. A fuzzy-logic-based approach to qualitative modeling. *IEEE Transactions on Fuzzy Systems*, 1: 7-33, 1993.

Y. Tamura, K. Suda, A. Sasaki, Y. Iwatani, K. Fujii, R. Ishibashi and K. Hibi. Simultaneous measurements of wind speed profiles at two sites using Doppler sodars. *Journal of Wind Engineering and Industrial Aerodynamics*, 89(3-4): 325-335, 2001.

M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1: 211–244, 2001.

J. L. Torres, A. García, M. De Blas and A. De Francisco. Forecast of hourly average wind speed with ARMA models in Navarre (Spain). *Solar Energy*, 79(1): 65-77, 2005.

R. P. Walker and A. Swift. *Wind energy essentials: Societal, economic, and environmental Impacts*, 1st edition. Wiley. 2015.

S. Yang and E. McKeogh. LIDAR and SODAR measurements of wind speed and direction in upland terrain for wind energy purposes. *Remote Sensing*, 3(9):1871-1901, 2011.

Fuzzy Clustering Algorithm applied to the Radio Frequency Signals Prediction

Paulo Tibúrcio Pereira Glaucio Lopes Ramos

GAPEA - Antennas and Propagation Research Group, Federal University of São João del-Rei (UFSJ), Brazil,
{paulotiburciop,glopesr}@gmail.com

Abstract

In this work the Fuzzy Clustering technique was used to perform radio frequency signals prediction. This technique was used with georeferencing maps of topography and morphology for prediction of radio frequency power levels, at the region of Viçosa - MG. The performance of this method was evaluated through tests of propagation and mapping for a 879.660 MHz signal, used in cellular mobile telephony. This method of prediction showed excellent results in comparison with measurements of RF power levels with a success rate greater than the classical models of Okumura-Hata and Walfisch-Ikegami. Consequently, this method can be very useful to the telecommunications companies when making the RF cellular coverage prediction.

Keywords: RF prediction, fuzzy clustering, mobile telephony

1 Introduction

The electromagnetic wave propagation between transmitter and receiver antennas has its characteristics fundamentally defined by the transmission medium between them. The radio signal has to propagate with low distortion and the received power must be adequately above the noise level in order to be correctly decoded.

To analyse Radio Frequency (RF) signals we must consider the electromagnetic waves, but also the topography and morphology of the terrain and, in some cases, the meteorological, ionospheric and spacial waves conditions. If these RF waves propagate in free space conditions, with no occurrence of reflection, diffraction, refraction, absorption attenuation, we would have ideal conditions to obtain the signal prediction.

In fact, the parameters of the medium where the electromagnetic waves propagate are strongly dependent on the specific propagation area, as forests, deserts, lakes, mountains, buildings, cities, and they also frequently vary due to atmospheric conditions, as temperature, pressure, humidity and noise.

2 Classical RF Prediction

The power signal strength, for a specific frequency, can be determined using (1):

$$P_r = P_t - L_p, \tag{1}$$

where P_r is the received power signal [dBm], P_t is Effective Isotropic Radiated Power (EIRP) [dBm] and L_p computes the total losses between the transmitter and receiver [dB].

These power losses can vary significantly due to the environments and atmospheric conditions previously described. The most usual way to compute the diffraction power losses are the Fresnel-Kirchoff equations and classical propagation models are also usually implemented in RF software predictions, as: Okumura-Hata, Walfisch-Ikegami Lee, COST 231 Hata, and COST 231 Walfish-Ikegami (Parsons, 2000; Hata, 1980; Walfisch and Bertoni, 1988; Lee, 1980; Ikegami et al., 1984). The Okumura-Hata model is applied to urban and suburbans environments. The signal strength attenuation is obtained by (2):

$$\begin{aligned} L_p = 69.55 + 26.16 log f_c - 13.82 log h_t - a(h_r) \\ + (44.9 - 6.55 log h_t) log d, \end{aligned} \tag{2}$$

where L_p is the signal attenuation [dB], f_c is the frequency [150-1500 MHz], d is the distance between the base and mobile stations [1-20 km], h_t is the effective height of the base station [30-200 m] and $a(h_r)$ is the correction factor for mobile antenna height, as (3),

$$a(h_r) = (1.1 log f_c 0.7) h_r (1.56 log f_c 0.8), \tag{3}$$

where h_r is the effective height of the mobile station [1-10 m].

The Walfisch-Ikegami model is used to predict the received signal in urban and dense urban environments. The signal attenuation when in a line-of-sight (LOS) situation is estimated by (4):

$$L_p(LOS) = 42.6 + 20 log f_c + 26 log d. \tag{4}$$

When in a non-line-of-sight condition (NLOS), the attenuation is predicted by (5):

$$\begin{aligned} L_p(NLOS) = 32.4 + 20 log f_c + 20 log d \\ + L_{diff} + L_{mult}, \end{aligned} \tag{5}$$

where f_c is the frequency [MHz], d is the distance [km], L_{diff} accounts for the diffraction losses on rooftops and L_{mult} accounts for the multiple diffraction in multiple buildings (Ikegami et al., 1984; Walfisch and Bertoni, 1988).

The coefficients for the equations of these classical prediction models are based in experimental data and statistical analysis. Some problems arises due the fact that there are different propagation environments. The RF prediction in the traditional way, are carried out by adjusting the coefficients of the equations in order to adjust them to a particular region, aiming to decrease the prediction errors.

3 Fuzzy Clustering Prediction

The classic prediction models uses restricted and empirical approaches of the RF propagation problem in certain regions. As opposite, the Fuzzy Clustering (Besdek, 1981) prediction method does not use empirical equations for calculating the received RF power (Pereira, 2000).

Fuzzy technology has been used in various areas of knowledge, such as control, decision making, pattern recognition, prediction of time series and state estimation and also in RF prediction (Phaiboon, 2010; Pelusi et al., 2014, 2013; Pelusi, 2012). It circumvents the resolution of complex differential equations with multiple variables applying the fuzzy rules and the compositional inference.

To make RF predictions it was chosen the ViÃğosa/MG region, which has an irregular topography and diverse morphology, thus promoting the emergence of various phenomena that influence the propagation of RF signals. It was collected signal sample levels at a cellular frequency of 879.660 MHz. Some of these samples were used for training and others to test the success rate. In this work it was used as input variables for the Fuzzy Clustering processing (Besdek, 1981), some RF measured samples performed in the field and georeferenced maps of topography and morphology (Pereira, 2000).

The proposal of using the fuzzy method was to circumvent the difficult in the modeling process of the propagation physical problems, featuring an unknown RF propagation environment from a group of known measurements, providing continuous mapping of RF coverage throughout a region. The method of RF signal strength prediction using Fuzzy Clustering and mapping is illustrated in Fig. 1 (Pereira, 2000).

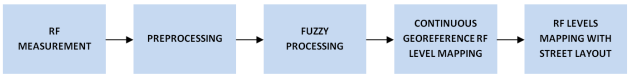


Figure 1. Fuzzy Clustering RF prediction technique (Pereira, 2000).

The first step of the method consists in measuring the RF signal. This was done through an automatic sampling process for the RF levels in some points of interest in the region under study. These measurements were performed using an RF receiver coupled to a GPS within a moving vehicle, thereby linking each measured RF signal level with its longitude and latitude coordinates. The equipment used was the HP E7474A equipped with the RF receiver model E6452A controlled by the software Viper. This system was developed by Agilent and adjusted to perform RF power measurements by distance, every 10 meters of vehicle movements.

In the second step it was carried out the pre-processing, aiming the data standardization and formatting. Initially it was calculated the distances, in relation to the transmitting antenna, of the points at which the RF power measurements were carried out.

For each point it was associated its information of the altitude and the region type, extracted from georeferenced maps of topography and morphology, as shown in Figs. 2 and 3. For each morphological type it was associated a numeric value, as follows: Low vegetation (field) = 1, rural area (plantation) = 2, scarce buildings = 3, airport = 4, forest (low vegetation bush) = 5, suburban area = 6, water (fresh-water) = 7, urban area = 8 and forest (tree) = 9. The altimetry and morphology databases were in the terrestrial ellipsoid model DATUM SAD69, cylindrical projection LAT/LON and had the planimetric resolution of 30 meters.

The use of databases with low planimetric resolution may compromise the results of RF predictions. For the RF signals predictions with higher frequencies, where the wavelength is smaller, it will be required higher resolution databases in order to better characterize the propagation environment and thus preserve the RF quality prediction.

All these data form a numerical table, so that each measurement point has now information on: longitude, latitude, distance, altitude, region type and RF level. These are the input variables for the Fuzzy Clustering processing.

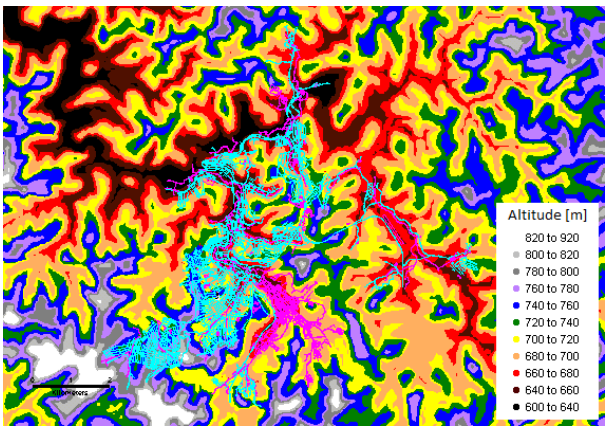


Figure 2. Georeferenced elevation contours map.

The variables previously obtained were used to train the Fuzzy Clustering process (Besdek, 1981), illustrated in Figs. 4 and 5, carried out in the two following distinct

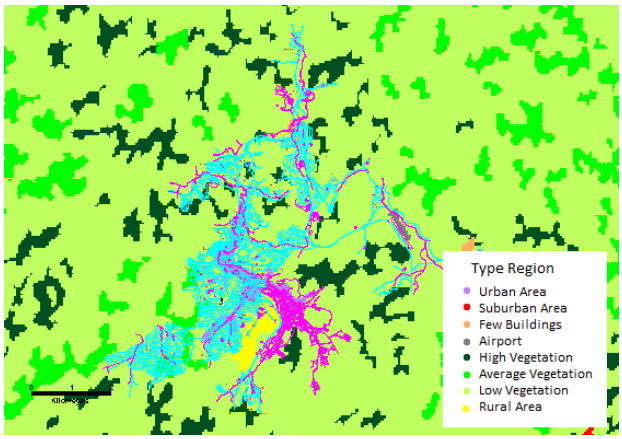


Figure 3. Georeferenced type region map.

steps.

In the first step, it was setted the number of fuzzy rules and their properties using the Grouping Estimation method that performs the following algorithm (Chiu, 1994; Yager and Filev, 1994):

1. A data vector with many neighboring data vectors have a high potential value;
2. The vector data with the greatest potential was the first assembly center;
3. An amount of potential from each data vector was subtracted as a function of distance from the first assembly center. The data vectors near from the first assembly center will have very limited potential, and therefore should not be selected as next assembly center;
4. It was selected the data vector with the largest remaining potential, as a second assembly center;
5. The number of clusters centroids was determined by (6):

$$P_k^* < \varepsilon P_1^*, \tag{6}$$

where P_k^* is the potential value of the $k-th$ cluster center, ε is a small decimal number between 0.15 and 0.50, and P_1^* is the potential value of the first assembly center.

In the second step, it was performed an optimization of the resulting rules. The Fuzzy Model Identification method was used to convert the parameters optimization of the resulting equations using a Linear Least Squares Estimation, as follows:

1. x_i^* are considered cluster centers, and will be the fuzzy rules that describe the system behavior;
2. x_i^* are decomposed in: y_i^* (input variables) and z_i^* (output variables). Being y the input vector, the value of relevance of the rule i is defined as (7):

$$\mu_i = e^{-\alpha \left\| y - y_i^* \right\|^2}, \tag{7}$$

where α is a constant. The output vector, z is calculated by (8):

$$z = \frac{\sum_{i=1}^{c} \mu_i z_i^*}{\sum_{i=1}^{c} \mu_i}, \tag{8}$$

z_i^* is considered a linear function of the input variables, as in the Takagi-Sugeno model (Sugeno, 1985):

$$z_i^* = G_i y + h_i, \tag{9}$$

where G_i is a constant matrix $(m-n)\cdot n$, and h_i is a constant column vector with $m-n$ elements. To convert the parameter optimization problem of (9) in a Linear Least Squares Estimation problem, it was defined:

$$\rho_i = \frac{\mu_i}{\sum_{i=1}^{c} \mu_j}. \tag{10}$$

Equation (8) can be rewritten as (11)

$$z = \sum_{i=1}^{c} \rho_i z_i^* = \sum_{i=1}^{c} \rho_i (G_i y + h_i), \tag{11}$$

or (12)

$$z^T = \begin{bmatrix} \rho_1 y^T & \rho_1 & \dots & \rho_c y^T & \rho_c \end{bmatrix} \begin{bmatrix} G_1^T \\ h_1^T \\ \vdots \\ G_c^T \\ h_c^T \end{bmatrix}, \tag{12}$$

where z^T and y^T are line vectors. With a collections of n input data points $y1, y2, ..., yn$ the resultant output collection is (13):

$$\begin{bmatrix} z_1^T \\ \vdots \\ z_n^T \end{bmatrix} = \begin{bmatrix} \rho_{1,1} y_1^T & \rho_{1,1} & \dots & \rho_{c,1} y_1^T & \rho_{c,1} \\ \rho_{1,n} y_n^T & \rho_{1,n} & \dots & \rho_{c,n} y_n^T & \rho_{c,n} \end{bmatrix} \begin{bmatrix} G_1^T \\ h_1^T \\ \vdots \\ G_c^T \\ h_c^T, \end{bmatrix} \tag{13}$$

where $\rho_{i,j}$ denotes ρ_i valued in y_j.

The first matrix on the right side of (13) is constant, while the second contains all the parameters to be optimized. To minimize the squared error between the output model and the training data, it was applied to the problem the Linear Least Squares Estimation technique (Chiu, 1994).

In the Least Squares Estimation process, (13) was placed in the form (14):

$$AX = B, \tag{14}$$

where B is the output data matrix, A is a constant matrix and X is the matrix of the parameters to be estimated.

The pseudo-inverse solution that minimizes

$$\|AX - B\|^2 \tag{15}$$

is given by (16):

$$X = (A^T A)^{-1} A^T B. \tag{16}$$

The $(A^T A)^{-1}$ calculation is computationally time consuming when $A^T A$ is a large array. The size of $A^T A$ is given by $c(n+1) \cdot c(n+1)$. Numerical problems also arise when $A^T A$ is almost singular. To solve these problems it was used the Recursive Least Squares method and X is obtained via the iterative formula (17) (Chiu, 1994):

$$X_{i+1} = X_i + S_{i+1} a_{i+1} \left(b_{i+1}^T - a_{i+1}^T X_i \right) \tag{17}$$

and (18)

$$S_{i+1} = S_i - \frac{S_i a_{i+1} a_{i+1}^T S_i}{1 + a_{i+1}^T a_{i+1}}, \tag{18}$$

where X_i is the X value estimated at $i-th$ iteration, S_i is a covariance matrix $c(N+1) \cdot C(N+1)$, a_i^T is the $i-th$ row vector of A and b_i^T is the $i-th$ vector B. The Least Squares Estimation of X matches the X_n value.

After the training was performed the Fuzzy Clustering activation processing was activated with the Longitude, Latitude, Distance, Altitude and Region Type data. The training data are the blue crosses and the activation data are the green crosses, as illustrated in Fig. 6. The activation grid was obtained from topography and morphology georeferenced maps. The spacing between grid points is the planimetric resolution of the maps. The results of this step are the RF levels at each point of the grid, forming a continuous mapping (Pereira, 2000).

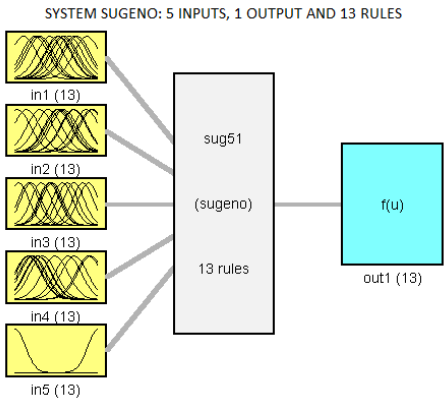


Figure 4. Processing Fuzzy Diagrams Blocks.

In the last step it was carried out the integration of the street map with the continuous mapping of RF levels. The RF levels were related with colors. These colors were superimposed on the streets in the region under study in order to facilitate the RF level preview of the important

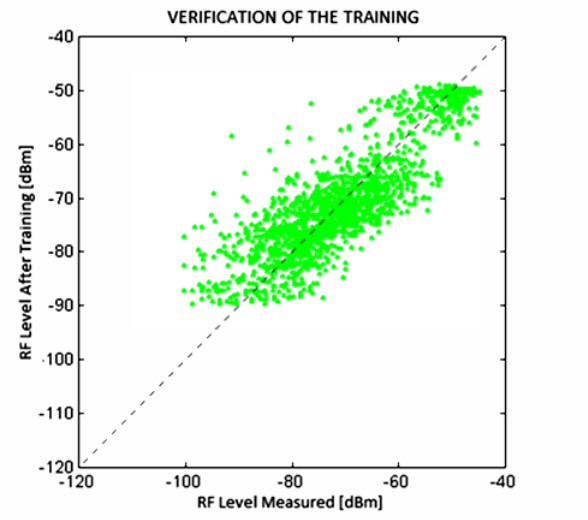


Figure 5. Verification Training.

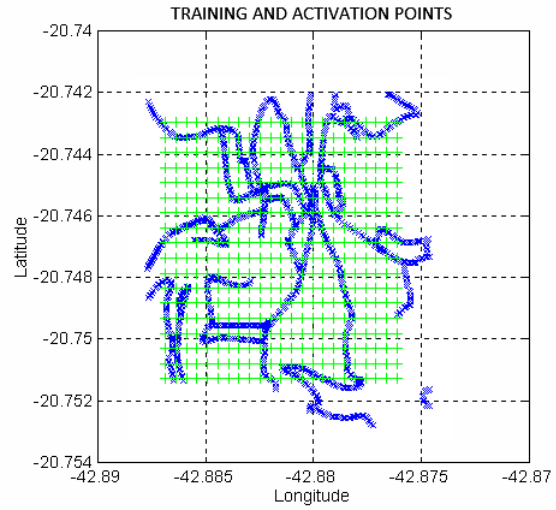


Figure 6. Training and Activation Points.

points. As a result, it was obtained the RF coverage map for the study area as in Fig. 7.

4 Analysis

The results of the RF Fuzzy Clustering Prediction method were compared with the results of the conventional prediction methods. To this end, it was used two measured data sets that were not used in the fuzzy prediction process. These two measured data sets were compared with the data obtained from the predictions. The mean absolute errors from the realized prediction methods can be seen in Table 1.

In Table 1 it is presented the comparison between the hit rate using the Fuzzy Clustering Prediction method and empirical techniques. The Fuzzy Clustering Prediction method was applied in three different scenarios: 3 (Longitude, Latitude and Distance), 4 (Longitude, Latitude, Distance and Altitude) and 5 (Longitude, Latitude, Distance,

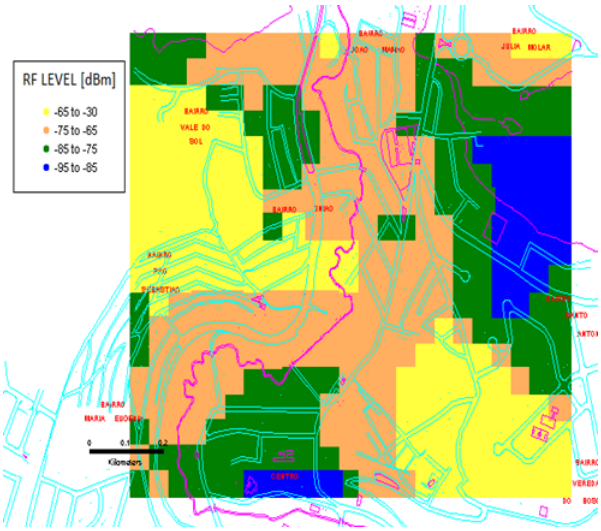


Figure 7. RF prediction with the Street Map.

Altitude and Region Type) input variables. It is evident the superiority of Fuzzy Clustering method in relation to the Okumura-Hata and Walsfish Ikegami classical methods.

Table 1. Arithmetic mean of the absolute errors of predictions

RF Prediction Method	Data Set 1 [dB]	Data Set 2 [dB]
Okumura-Hata	9.8596	18.0591
Walfisch-Ikegami	16.7718	10.8308
Fuzzy Clustering 3 input variables	5.9252	10.8028
Fuzzy Clustering 4 input variables	5.9537	9.5681
Fuzzy Clustering 5 input variables	5.9645	8.2682

5 Conclusions

The use of altimetry, morphology database and RF georeferenced sample levels can characterize the propagation environment and in conjunction with the fuzzy logic processing provides small errors in RF predictions.

This method of prediction showed excellent results in comparison with the actual RF power levels, with a success rate greater than some classical models as Okumura-Hata and Walfisch-Ikegami.

Using the prediction method developed in this work it was generated a continuous RF coverage mapping integrated to the street layout of the study area, becoming an application of fuzzy logic in the telecommunications area.

The method described in this paper can be used by wireless telecommunications companies to perform RF predictions quickly and accurately. It also includes carries operating with different frequencies or other regions with different characteristics.

The Fuzzy Clustering RF prediction process, described in this work, together with high resolution databases may assist in the deployment of future mobile telecommunications networks.

Acknowledgment

This work has been supported by the Brazilian agency CAPES/PROCAD 068419/2014-01.

References

J. C. Besdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.

S. Chiu. Fuzzy model identification based on cluster estimation. *Journal of Intelligent & Fuzzy Systems*, 2(3), 1994.

M. Hata. Empirical formula for propagation loss in land mobile radio services. *IEEE Transactions on Vehicular Technology*, 29(3):317–325, August 1980.

F. Ikegami, S. Yoshida, T. Takeuchi, and M. Umehira. Propagation factors controlling mean field strength on urban streets. *IEEE Trans. On Antennas and Propagation*, 32(8):822 – 829, August 1984.

W. C. Y. Lee. Studies of base-station antenna height effects on mobile radio. *IEEE Trans. on Vehicular Technology*, 29(2): 252–260, May 1980.

J. D. Parsons. *Mobile Radio Propagation Channel*. Wiley, 2000.

D. Pelusi. PID and intelligent controllers for optimal timing performances of industrial actuators. *International Journal of Simulation: Systems, Science and Technology*, 13(2):65–71, 2012.

D. Pelusi, L. Vazquez, D. Diaz, and R. Mascella. Fuzzy algorithm control effectiveness on drum boiler simulated dynamics. *36th International Conference on Telecommunications and Signal Processing*, pages 272–276, 2013.

D. Pelusi, M. Tivegna, and P. Ippoliti. Intelligent algorithms for trading the euro-dollar in the foreign exchange market? *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pages 243–252., 2014.

P.T. Pereira. Predição de sinais de radiofreqüência utilizando lógica fuzzy. Master's thesis, CEFET/MG, 2000.

Supachai Phaiboon. RF Macro-cell Prediction Using Fuzzy Logic: Case study in Bangkok City - Thailand. In *Proceedings of the 4th International Conference on Communications and Information Technology*, CIT'10, pages 105–109, Stevens Point, Wisconsin, USA, 2010. World Scientific and Engineering Academy and Society (WSEAS). ISBN 978-960-474-207-3. URL `http://dl.acm.org/citation.cfm?id=1864098.1864117`.

M. Sugeno. *Industrial applications of fuzzy control*. Elsevier Science Pub. Co., 1985.

J. Walfisch and H. L. Bertoni. A theoretical model of UHF propagation in urban environments. *IEEE Transactions on Antennas and Propagation*, 36(12):1788–1796, December 1988.

R. Yager and D. Filev. Generation of fuzzy rules by mountain clustering. *Journal of Intelligent & Fuzzy Systems*, 2(3):209–219, 1994.

Single Swarm and Simple Multi-Swarm PSO Comparison

Michal Pluhacek[1], Roman Senkerik[1], Adam Viktorin[1] and Ivan Zelinka[2]

[1]Faculty of Applied informatics
Tomas Bata University in Zlin, Nam T.G. Masaryka 5555, 760 01Zlin, Czech Republic
{pluhacek,senkerik,aviktorin}@fai.utb.cz

[2]Department of Computer Science, Faculty of Electrical Engineering and Computer Science
VSB-TUO, 17.listopadu 15, 708 33 Ostrava–Poruba, Czech Republic
ivan.zelinka@vsb.cz

Abstract

This paper presents a comparative study of the original single swarm PSO with linear decreasing inertia weight and basic multi-swarm variant. The alternative population topology (multi-swarms) is a topic that is getting increased attention from the research community in recent years. We present evidence that it might be very beneficial to divide the population into two sub-swarms with partially restricted communication. The size of the sub-swarms is chosen with respect to a previously published study on this topic. The scaling of the methods is compared in three different dimensional settings. The results are statistically evaluated and discussed. The paper concludes with proposals for future research.

Keywords: particle swarm optimization, swarm intelligence, multi-swarm, topology

1 Introduction

In the past decades, the Particle swarm optimization (PSO) algorithm has established itself as a very efficient global optimizer. The original design (Kennedy and Eberhart, 1995; Kennedy, 1997) has been regularly modified (Shi and Eberhart, 1998; Engelbrecht 2010) and studied in great detail (van den Bergh and Engelbrecht 2006).

One of the most popular approaches for PSO modification is the multi-swarm (Liang and Suganthan, 2005; Ostadrahimi et al. 2012; Solomon et al., 2011; Dor, 2012; Liu et al., 2011) approach. In the multi-swarm approaches, the population is divided into multiple sub-populations (sub-swarms) with different levels of communication. The benefit of such approach is in that the population can maintain divergence, search multiple promising regions and partially converge into multiple optima.

In (García-Nieto and Alba, 2012) the optimal swarm (sub-swarm) size is discussed in great detail. It is proposed that six particles per swarm might be the optimal number for PSO based algorithms.

In this study, we build on those findings and compare the performance of single swarm PSO with a PSO with two sub-swarms that consist of six particles each.

The rest of the paper is structured as follows: In the following section, the PSO algorithm is described. In the next section, the proposed simple multi-swarm algorithm is introduced. Used test functions in this work are presented in the following section. The experiment is set up and results presented in the following sections. The paper concludes with the discussion of the results.

2 Particle Swarm Optimization (PSO)

The PSO algorithm is inspired by the natural swarm behavior of animals (such as birds and fish). It was firstly introduced by Eberhart and Kennedy (1995). Soon the PSO became a popular method for global optimization. Each particle in the population represents a possible solution of the optimization problem which is defined by the cost function (CF). In each iteration of the algorithm, a new location (combination of CF parameters) of the particle is calculated based on its previous location and velocity vector (velocity vector contains particle velocity for each dimension).

According to the method of selection of the swarm or subswarm for best solution information spreading, the PSO algorithms are noted as global PSO (GPSO) or local PSO (LPSO). Within this research, the PSO algorithm with global topology was utilized. The velocity calculation formula is given by

$$v_{ij}^{t+1} = w \cdot v_{ij}^{t} + c_1 \cdot Rand \cdot (pBest_{ij} - x_{ij}^{t}) + c_2 \cdot Rand \cdot (gBest_{j} - x_{ij}^{t}) \quad (1)$$

where
v_{ij}^{t+1} - New velocity of the ith particle in iteration t+1. (component j of the dimension D).
w – Inertia weight value.
v_{ij}^{t} - Current velocity of the ith particle in iteration t. (component j of the dimension D).
c_1, c_2 - Acceleration constants.
$pBest_{ij}$ – Local (personal) best solution found by the ith particle. (component j of the dimension D).

$gBest_j$ - Best solution found in a population. (component j of the dimension D).
x_{ij}^t - Current position of the ith particle (component j of the dimension D) in iteration t.
Rand – Pseudo random number, interval (0, 1).

The maximum velocity was limited to 20% of the dimension range as it is usual. The new position of each particle is then given by:

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{2}$$

where x_i^{t+1} is the new particle position

Finally, the linear decreasing inertia weight ((Shi and Eberhart, 1998) is used in the typically referred PSO design that was used in this study. The dynamic inertia weight is meant to slow the particles over time thus to improve the local search capability in the later phase of the optimization. The inertia weight has two control parameters w_{start} and w_{end}. A new w for each iteration is given by (3), where t stands for current iteration number and n stands for the total number of iterations. The values used in this study were w_{start} = 0.9 and w_{end}=0.4.

$$w = w_{start} - \frac{((w_{start} - w_{end}) \cdot t)}{n} \tag{3}$$

3 Proposed multi-swarm model

In this initial study, a simple multi-swarm PSO was utilized. In the proposed approach the population is divided into two subswarms. Each subswarm utilizes different gBest for particle movement. After every 100 iterations, the values of *gBest* of swarm 1 and *gBest* of swarm 2 are compared, and the gBest with better value is copied as a gBest for both swarms.

Using this simple method, the population-wide communication is restricted to periodical time-windows.

4 Test Functions

Within this initial research, four well-known and frequently used benchmark functions were utilized. The optimum value is 0 for all functions. Four benchmark functions are used in the comparisons.

Sphere function

$$f_1(x) = \sum_{i=1}^{D} x_i^2 \tag{4}$$

where D - dimension
Search Range: $[-100,100]^D$
Glob. Opt. Pos.: $[0]^D$

Schwefel´s function

$$f_5(x) = 418.9829 \cdot \mathrm{D} \sum_{i=1}^{D} -x_i \sin(\sqrt{|x|}) \tag{5}$$

Search Range: $[-500,500]^D$
Glob. Opt. Pos.: $[420.96]^D$

***Rastrigin´s* function**

$$f_6(x) = \sum_{i=1}^{D} [x_i^2 - 10\cos(2\pi x_i) + 10] \tag{6}$$

Search Range: $[-5.12,5.12]^D$
Glob. Opt. Pos.: $[0]^D$

Rosenbrock´s function

$$f_3(x) = \sum_{i=1}^{D-1} [100(x_i^2 - x_{i+1})^2 + (1 - x_i)^2] \tag{7}$$

Search Range: $[-10,10]^D$
Glob. Opt. Pos.: $[0]^D$

5 Experiment setup

Two variants of the PSO algorithm were compared in this study with 3 different dimension settings. The single-swarm PSO with linear decreasing inertia weight (as described in section 2) noted further PSO. The multi-swarm variant proposed in section 3 is noted further PSO Multi.
The control parameters were set as follows:

Population size: 12 (PSO) 6+6 (PSO Multi)
Iterations: 2500
v_{max}: 0.2
w_{start}: 0.9
w_{end}: 0.4
c_1,c_2 = 2
Dim = 5, 10, 20

6 Results

In this section results obtained in the experiments with both previously described PSO variants are presented and compared. In table 1 – 12 the statistical overview of results from 30 repeated runs of each algorithm are given alongside with the p-value of Wilcoxon signed-rank test with significance level 0.05. The best mean results are highlighted by bold numbers.

Table 1. Results - Sphere function: dim: 5.

Sphere dim: 5	PSO	PSO Multi	p-value
Mean CF Value:	2.051E-14	**1.136E-88**	1.825E-06
Std. Dev.:	1.123E-13	5.367E-88	
CF Value Median:	6.307E-31	8.659E-93	
Max, CF Value:	6.152E-13	2.928E-87	
Min, CF Value:	1.914E-45	4.884E-98	

Table 2. Results - Schwefel function: dim: 5.

Schwefel dim: 5	PSO	PSO Multi	p-value
Mean CF Value:	4.106E+02	**2.645E+02**	9.280E-04
Std. Dev.:	1.697E+02	1.344E+02	
CF Value Median:	4.540E+02	2.369E+02	
Max, CF Value:	7.896E+02	5.922E+02	
Min, CF Value:	1.184E+02	6.364E-05	

Table 3. Results - Rastrigin function: dim: 5.

Rastrigin dim: 5	PSO	PSO Multi	p-value
Mean CF Value:	2.056E+00	**3.980E-01**	4.072E-06
Std. Dev.:	1.522E+00	5.604E-01	
CF Value Median:	1.990E+00	0.000E+00	
Max, CF Value:	5.970E+00	1.990E+00	
Min, CF Value:	2.842E-14	0.000E+00	

Table 4. Results - Rosenbrock function: dim: 5.

Rosenbrock dim: 5	PSO	PSO Multi	p-value
Mean CF Value:	2.621E-01	**1.389E-01**	3.732E-04
Std. Dev.:	9.973E-01	4.911E-02	
CF Value Median:	1.344E-19	1.273E-01	
Max, CF Value:	3.931E+00	2.496E-01	
Min, CF Value:	0.000E+00	6.849E-02	

Table 5. Results - Sphere function: dim: 10.

Sphere dim: 10	PSO	PSO Multi	p-value
Mean CF Value:	4.586E-04	**2.635E-41**	1.825E-06
Std. Dev.:	4.364E-04	6.049E-41	
CF Value Median:	2.739E-04	1.185E-42	
Max, CF Value:	1.737E-03	2.155E-40	
Min, CF Value:	3.916E-06	1.328E-45	

Table 6. Results - Schwefel function: dim: 10.

Schwefel dim: 10	PSO	PSO Multi	p-value
Mean CF Value:	1.443E+03	**8.751E+02**	2.237E-06
Std. Dev.:	2.209E+02	2.307E+02	
CF Value Median:	1.423E+03	8.982E+02	
Max, CF Value:	2.057E+03	1.323E+03	
Min, CF Value:	1.052E+03	3.356E+02	

Table 7. Results - Rastrigin function: dim: 10.

Rastrigin dim: 10	PSO	PSO Multi	p-value
Mean CF Value:	9.356E+00	**4.676E+00**	3.561E-05
Std. Dev.:	2.472E+06	2.864E+00	
CF Value Median:	9.268E+00	3.980E+00	
Max, CF Value:	1.413E+01	1.492E+01	
Min, CF Value:	2.351E+00	9.950E-01	

Table 8. Results - Rosenbrock function: dim: 10.

Rosenbrock dim: 10	PSO	PSO Multi	p-value
Mean CF Value:	7.560E+00	**3.743E+00**	4.502E-06
Std. Dev.:	2.194E+06	9.581E-01	
CF Value Median:	7.805E+00	3.683E+00	
Max, CF Value:	1.519E+06	8.144E+00	
Min, CF Value:	2.619E+00	1.184E+06	

Table 9. Results - Sphere function: dim: 20.

Sphere dim: 20	PSO	PSO Multi	p-value
Mean CF Value:	6.930E-02	**8.433E-17**	1.825E-06
Std. Dev.:	4.192E-02	1.890E-16	
CF Value Median:	5.756E-02	1.003E-17	
Max, CF Value:	1.705E-01	8.765E-16	
Min, CF Value:	2.370E-02	8.788E-20	

Table 10. Results - Schwefel function: dim: 20.

Schwefel dim: 20	PSO	PSO Multi	p-value
Mean CF Value:	4.035E+03	**2.452E+03**	1.825E-06
Std. Dev.:	3.237E+02	3.954E+02	
CF Value Median:	3.996E+03	2.497E+03	
Max, CF Value:	4.566E+03	3.079E+03	
Min, CF Value:	3.335E+03	1.796E+03	

Table 11. Results - Rastrigin function: dim: 20.

Rastrigin dim: 20	PSO	PSO Multi	p-value
Mean CF Value:	3.283E+01	**1.983E+01**	5.478E-06
Std. Dev.:	8.882E+00	6.411E+00	
CF Value Median:	3.129E+01	1.841E+01	
Max, CF Value:	5.981E+01	3.482E+01	
Min, CF Value:	1.929E+01	7.960E+00	

Table 12. Results - Rosenbrock function: dim: 20.

Rosenbrock dim: 20	PSO	PSO Multi	p-value
Mean CF Value:	1.955E+01	**1.563E+01**	2.475E-06
Std. Dev.:	1.573E+00	1.700E+00	
CF Value Median:	1.963E+01	1.502E+01	
Max, CF Value:	2.337E+01	1.950E+01	
Min, CF Value:	1.583E+01	1.351E+01	

7 The scaling comparison

In this section, the comparison of scalabilities of the methods is presented. The mean results for three different dimensional settings (5, 10 and 20) are compared in Figures 1 – 4.

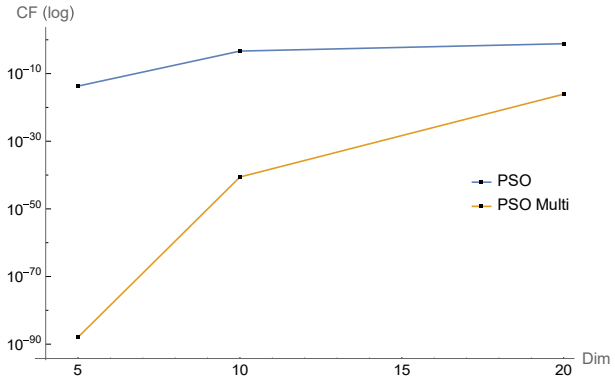


Figure 1. Scaling comparison – Sphere function.

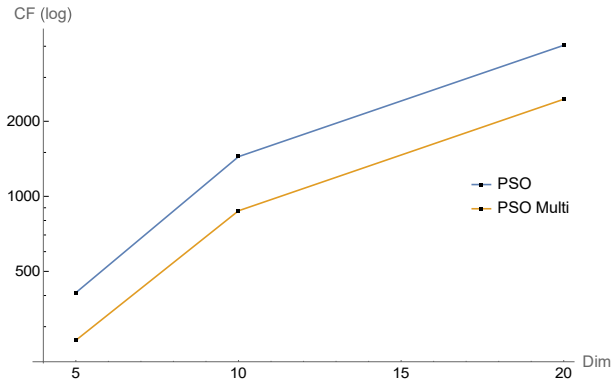


Figure 2. Scaling comparison – Schwefel´s function.

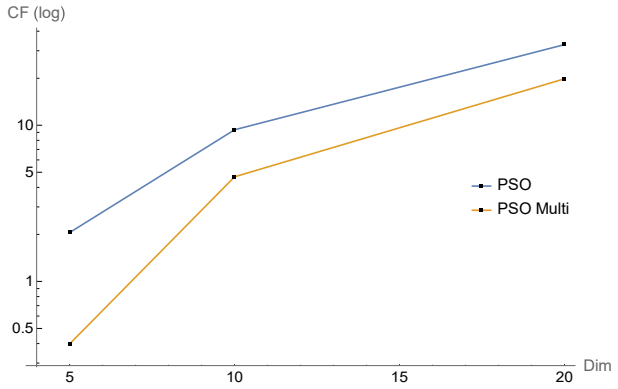


Figure 3. Scaling comparison – Rastrigin´s function.

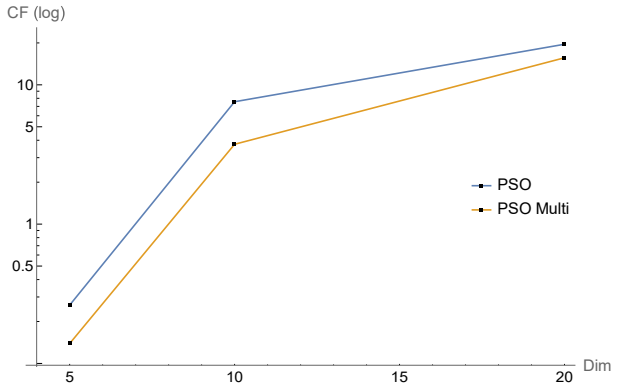


Figure 4. Scaling comparison – Rosenbrock´s function.

8 Results and discussion

According to data presented in Tables 1 – 12 the Multi-swarm variant performance is better for all four different test functions in all dimensional settings. The results are statistically significant according to the Wilcoxon signed-rank test. It is worth noting that the results are in many cases significantly better despite that both algorithms utilize the similar total number of particles and the number of iterations.

As is presented in the scalability comparison (Figures 1 – 4) the difference in performance is usually getting smaller with increasing dimensionality of the problem. The performance in higher dimensions will be addressed in future studies.

9 Conclusions

In this paper, a comparison of a single swarm and multi-swarm PSO was presented. Based on the previous works of (García-Nieto and Alba, 2012) the swarm size for the multi-swarm approach was set to 6 particles. It has been presented that such algorithm is capable of obtaining very good results on four typically used benchmark functions. Also, the performance was in all cases superior to the performance of single swarm PSO algorithm.

The results are surprisingly consistent, and this initial study will be followed by a more detailed study with

multiple population size settings and more test functions.

Based on the initial results it seems that using basic multi-swarm approach in opposite to a single swarm might be beneficial for PSO based algorithms. The future research will focus on multi-swarm approaches for advanced PSO modifications.

Acknowledgements

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme Project no. LO1303 (MSMT-7778/2014), further by the European Regional Development Fund under the Project CEBIA-Tech no. CZ.1.05/2.1.00/03.0089 and by Internal Grant Agency of Tomas Bata University under the Projects no. IGA/CebiaTech/2018/003. This work is also based upon support by COST (European Cooperation in Science & Technology) under Action CA15140, Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO), and Action IC1406, High-Performance Modelling and Simulation for Big Data Applications (cHiPSet). The work was further supported by resources of A.I.Lab at the Faculty of Applied Informatics, Tomas Bata University in Zlin (ailab.fai.utb.cz).

References

A. El Dor, M. Clerc, and P. Siarry. A multi-swarm PSO using charged particles in a partitioned search space for continuous optimization. *Computational Optimization and Applications*, 53(1): 271-295, 2012.

A. Engelbrecht. Heterogeneous particle swarm optimization. In *Proceedings of the 7th international conference on Swarm intelligence*. Berlin, Heidelberg: Springer-Verlag, pages 191–202, 2010.

J. García-Nieto, E. Alba. Why six informants is optimal in PSO. In *Proceedings of the 14th annual conference on Genetic and evolutionary computation,* pages 25-32, ACM. 2012.

J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of the IEEE International Conference on Neural Networks*, pages 1942–1948, 1995.

J. Kennedy. The particle swarm: social adaptation of knowledge. In *Proceedings of the IEEE International Conference on Evolutionary Computation*, pages 303–308, 1997.

J.-J. Liang and P.N. Suganthan. Dynamic multi-swarm particle swarm optimizer with local search. In *Evolutionary Computation, The 2005 IEEE Congress on*. IEEE, 2005. pages 522-528, 2005.

Y. Liu, G. Wang, H. Chen, H. Dong, X. Zhu, and S. Wang. An improved particle swarm optimization for feature selection. *Journal of Bionic Engineering*, 8(2): 191-200, 2011.

L. Ostadrahimi, A. M.M iguel, and A. Abbas. Multi-reservoir operation rules: multi-swarm PSO-based optimization approach. *Water resources management*, 26(2): 407-427, 2012.

S. Solomon, P. Thulasiraman, and R. Thulasiram. Collaborative multi-swarm PSO for task matching using graphics processing units. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation,* pages 1563-1570, ACM. 2011.

Y. Shi and R. Eberhart. A modified particle swarm optimizer. In *Proceedings of the IEEE International Conference on Evolutionary Computation* (IEEE World Congress on - Computational Intelligence), pages 69–73. I. S. 1998.

F. van den Bergh and A.P. Engelbrecht. A study of particle swarm optimization particle trajectories. *Information Sciences,* 176(8): 937-971, 2006.

Flow Rate Estimation using Dynamic Artificial Neural Networks with Ultrasonic Level Measurements

Khim Chhantyal Minh Hoang Håkon Viumdal Saba Mylvaganam

Faculty of Technology, Natural Sciences, and Maritime Sciences, University College of Southeast Norway,
{khim.chhantyal,hakon.viumdal,saba.mylvaganam}@usn.no, m.hoang1304@gmail.com

Abstract

Accurate estimation of flow in drilling operations at inflow and outflow positions can lead to increased safety, optimized production and improved cost efficiency. In this paper, Dynamic Artificial Neural Network (DANN) is used to estimate the flow rate of non-Newtonian drilling fluids in an open channel Venturi-rig that may be used for estimating outflow. Flow in the Venturi-rig is estimated using ultrasonic level measurements. Simulation study looks into fully connected Recurrent Neural Network (RNN) with three different learning algorithms: Back Propagation Through Time (BPTT), Real-Time Recurrent Learning (RTRL) and Extended Kalman Filter (EKF). The simulation results show that BPTT and EKF algorithms converge very quickly as compared to RTRL. However, RTRL gives more accurate results, is less complex and computationally fastest among these three algorithms. Hence, in the experimental study RTRL is chosen as the learning algorithm for implementing Dynamic Artificial Neural Network (DANN). DANN with RTRL learning algorithm is compared with Support Vector Regression (SVR) and static ANN models to assess their performance in estimating flow rates. The comparisons show that the proposed DANN is the most accurate model among three models as it uses previous inputs and outputs for the estimation of current output.

Keywords: drilling operations, open channel venturi flume, non-Newtonian fluid, flow rate estimation, ultrasonic level measurements, recurrent neural network, real-time recurrent learning

1 Introduction

In drilling operations, the drilling mud is circulated in a closed loop starting from the mud tank into the wellbore and back to the mud tank. The mud can be water-based, oil-based or gas-based and is circulated during the drilling operation, until the desired depth is reached. During circulation, the rheological properties of drilling mud have significant importance for the safe and efficient drilling operation. The viscosity, density, and flow rate of circulating mud play a vital role, in all the drilling operations. (Caenn et al., 2011)

In general, drilling muds are non-Newtonian in nature, and the viscosity of the mud along with other rheological properties govern the transport of rock cuttings while drilling. (Caenn et al., 2011)

The density or mud weight is mainly responsible for maintaining the pressure in the wellbore. Depending on the types of the drilling operation and the reservoir, the wellbore pressure or bottom-hole pressure (P_b) is limited within the pressure window given by formation pressure (P_f) and formation fracture pressure (P_{ff}). If the wellbore pressure is less than the formation pressure ($P_b < P_f$), the formation gasses and fluids will flow into the drilling mud, and is called *kick*. The occurrences of kick should be detected as early as possible during drilling operations. If the early kick detection is ignored or is not detected, it can lead to problems in maintaining the density of the mud and in the extreme case, it can result in blow-out of hydrocarbons on the rig, e.g. the Deepwater Horizon explosion, (Hauge and Øien, 2012). In the case of ($P_b > P_f$), the high pressure circulating fluids may enter the formation pores, causing fluid losses. If the wellbore pressure is further increased, beyond the formation fracture pressure ($P_b > P_{ff}$), the circulation fluid can fracture the formation and cause an increased fluid loss, often called lost circulation. The fluid loss will decrease the volume of the mud in the circulation loop and in the mud tank, and will affect the production, (Caenn et al., 2011).

A similar situation occurs frequently in geothermal drilling. In geothermal drilling, one of the costly problems is lost circulation. that occurs when drilling fluid is lost to the formation rather than returning to the surface, preferably intact. The management of lost circulation is important and requires the accurate measurement of drilling fluid flow rate both into and out of the well.

Reliable detection of unusual conditions can allow the use of low weight mud, efficient drilling, less formation damage, and lead to lower drilling costs. Delta flow method, i.e. calculating the difference between flows at inflow and outflow points of the circulation mud, is one of the best methods to detect kick and fluid loss, which uses the flow measurements before and after the wellbore, (Maus et al., 1979; Speers et al., 1987; Orban et al., 1987, 1988; Schafer et al., 1991; Kamyab et al., 2010). The difference in outflow and inflow measurements can be used as an indication of unusual conditions while drilling. If the flow rate before wellbore is less than the flow rate in the return line, then it can be considered as an indication of

early kick detection. Whereas, if the inflow is greater than the outflow, it is an early indication of fluid loss. In addition, the flow rate of circulating fluid will determine the transportation of rock cuttings. The flow velocity of the circulation mud is often maintained higher than the settling velocity of the rock cuttings for efficient transportation of cuttings. In addition to the delta flow method, other methods of early kick detection are discussed in (Kamyab et al., 2010; Mills et al., 2012; Ali et al., 2013; Patel et al., 2013; Vajargah et al., 2013).

In literature (Maus et al., 1979; Speers et al., 1987; Orban et al., 1987, 1988; Schafer et al., 1991; Kamyab et al., 2010), there are different systems for measuring delta flow. For inflow measurement, conventional pump stroke counter, rotatory pump speed counter, magnetic flow meter, Doppler ultrasonic flow meter or Coriolis mass flowmeter can be used. For outflow measurement, magnetic flowmeter, Doppler ultrasonic-based flowmeter, standard paddle meter, ultrasonic level meter, a prototype rolling float meter or open channel Venturi flowmeter can be used. The scenario of inflow and outflow measurement is completely different. For example, the inflow measurement can be carried out using Coriolis mass flow meter, more accurate but an expensive flowmeter. However, Coriolis mass flow meter is not suitable for outflow measurements as the returning mud contains solid rock cuttings, other formation particles, formation fluids and gases. An overview of different flowmeters based on reliability and accuracy is given in (Schafer et al., 1991). Based on this analysis, magnetic flowmeter or Doppler ultrasonic flowmeter are suitable for inflow measurements and prototype rolling float meters for outflow measurement. (Speers et al., 1987) presents the implementation of delta flow method by using magnetic flowmeters at inflow and outflow locations. The magnetic flowmeter is limited in applications to conductive fluids or to only water-based muds. In addition, magnetic flowmeters need some additional U-tube design in the return section. For lower flow velocity of circulating fluids, the rock cuttings will settle at the bottom of this U-tube. These problems are avoided in open channel return line, in which efficient rock cutting transportation and their easier separation from mud, (Orban et al., 1987, 1988).

This paper presents the outflow measurement based on open channel flow with a Venturi section. In an open channel flow, the upstream pressure relative to a reference level in the control section of the loop structure can be used to estimate the flow rate, (White, 2002). The control section used in the flow loop is the Venturi flume. The flow measurement is based on an extension of the application of the well-known Venturi principle, to flow of fluid in an open channel, (Skorpik, 2013). The constriction in the Venturi section results in the transition of flow from subcritical to supercritical flow in the vicinity of the throat, (Frenzel, 2011). For sufficiently long throat, the critical condition occurs in the throat, giving the critical depth, (Geratebau, 2013). The level of the fluid in upstream is measured as the critical depth is identified. The level can be measured using ultrasonic or RADAR level sensors and flow rate can be calculated as a function of measured level.

To study the possibility of using Venturi flume in estimating flow rate, a flow loop (i.e. Venturi rig) is available in University College of Southeast Norway (USN), Porsgrunn, Norway. For this Venturi rig, the CFD simulation study of open channel flow measurement is investigated in (Berg et al., 2013). The numerical algorithm using Saint Venant equation is presented in (Agu and Lie, 2014a,b). However, the developed numerical model is not applicable for real-time monitoring and controlling purpose due to the high computational cost. The study presented in (Chhantyal et al., 2016b) shows the successful implementation of static Artificial Neural Network (ANN) and Support Vector Regression (SVR) techniques for flow measurement in the test loop. The present study is a continuation, where, Dynamic Artificial Neural Networks (DANN) are investigated and implemented in the software used in running, monitoring and controlling the flow loop.

In the following sections, the simulation study of fully connected Recurrent Neural Network (RNN) with three different learning algorithms for estimating the flow rate of the non-Newtonian liquids is presented. Finally, the experimental results of flow rate estimation using RNN, ANN and SVR are discussed.

2 Dynamic Artificial Neural Network

ANN can be of the static or dynamic type. Static ANN or feedforward ANN type uses current inputs and current outputs whereas, DANN uses current and previous inputs and outputs for modeling purpose. Further, DANN can be partially connected RNN or fully connected RNN based on the feedback loops. Fully connected RNNs have self-feedback loops, and partially connected RNNs does not have self-feedback loops (Dijk, 1999). Some useful details of the MATLAB Toolbox DANN are given in the companion paper in this volume (Chhantyal et al., 2016a).

The delta flow measurement discussed in Section 1 is a dynamic problem, where the previous information about the kick detection and fluid loss is important for the current measurement. Therefore, fully connected RNN is used for modeling, the estimation of the flow rate being based on the level measurements in the open channel Venturi flow loop. For the estimation of the flow rate, three different learning algorithms are used. These algorithms are presented here briefly.

2.1 Back Propagation Through Time (BPTT)

BPTT is an extension of gradient-based back propagation algorithm that is used in static ANN. The idea in BPTT is to unfold the RNN architecture into feedforward ANN architecture in an arbitrary number of time steps or folds. These folds make the error to propagate even further in time, so it is called back propagation through time. However, the number of folds are usually low to avoid deep network and this approach is called is often called trun-

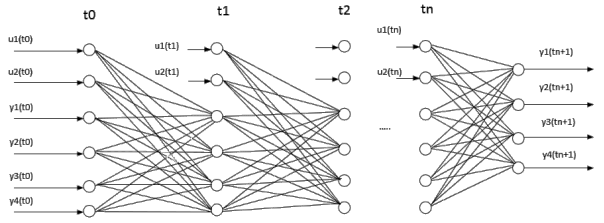


Figure 1. A general architecture for Back Propagation Through Time (BPTT) learning algorithm with N number of neurons and n numbers of foldings.

cated BPTT. In general, recurrent weights are simply duplicated over the folds while unfolding, (Boden, 2001). The basic BPTT architecture is shown in Figure 1. The computational complexity of BPTT is of order $O(N^2)$ and the storage requirement is of order $O(N^2)$, where N being number of neurons and n is the arbitrary number of folds. The drawbacks of BPTT are; it is an offline learning algorithm and requires large memory to store state information at different folds (Williams, 1992).

2.2 Real Time Recurrent Learning (RTRL)

RTRL is one of the most used real-time learning algorithms for RNN. In RTRL, the gradients at timet are computed based on the gradients at previous time steps. The gradient information is propagated in time (Mak et al., 1999; Mandic and Chambers, 2000; Budik, 2006). The basic RTRL architecture is shown in Figure 2. The connections with blue color are the additional self-feedback and feedback connections, which is not included in static ANN. These additional connections make the network get previous input values and output values and consider them as additional internal inputs in the current time. By doing this, a network can work dynamically. However, RTRL algorithm suffers from slow convergence, which is typical for all gradient-based algorithms. Mandic and Chambers, (Mandic and Chambers, 2000) has presented an RTRL-based learning algorithm with an adaptive learning rate that can improve the convergence performance. RTRL further suffers from the large computational complexity of the order of $O(N^4)$ and even critically with a storage requirement of the order of $O(N^3)$, (Williams, 1992).

2.3 Extended Kalman Filter Learning (EKF)

EKF is a recursive algorithm that computes state estimations based on the previous state information at the current time, (Kim, 2011). EKF can be used as a supervised on-line learning algorithm to determine the weights of an RNN. In EKF learning algorithm, the state vector consists of weights and the locally induced outputs of each neuron in the network. Regarding convergence to a solution, EKF is very fast compared to BPTT and RTRL. The order of computational complexity for EKF is same as RTRL, $O(N^4)$, and the storage requirement increases to the order of $O(N^4)$ for EKF. The RTRL algorithm is identical to the simplified EKF algorithm, and the architecture is the same, (Williams, 1992).

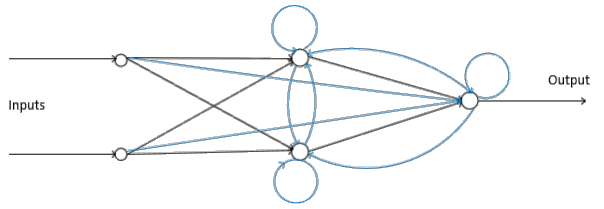


Figure 2. A general architecture for Recurrent Neural Network (RNN) with self-feedback and feedback loops from neurons.

3 Experimental Set-up

To develop RNN models, model-drilling fluid is circulated in the flow loop. The circulated fluid is visco-plastic in nature with the fluid properties of density at 1136 kg/m3 and a viscosity ranging from $23-180$ [centipoise] for the $500-1$ [s^{-1}] shear rate. Figure 3 shows the open channel section of flow rig with a Venturi constriction and three ultrasonic level sensors. The mass flow measurement is performed using Coriolis mass flow meter and is considered as a reference for RNN models. Recent study shows that the level measurements at the throat (LT-18), the level of the downstream (LT-17) and the level of the upstream (LT-15) are highly correlated to flow rate (Chhantyal et al., 2016b). Therefore, these variables are considered for modeling and are given in Table 1 and some concurrent measurements from these three ultrasonic sensors are shown in Figure 4, along with simultaneous measurements of flow from a Coriolis meter.

For the mass flow rate range of 250-500 [kg/min], 1800 data samples for each variable are measured. The data samples are normalized in the range of (0-1). Out of 1800 normalized data samples, 70%, 15% and 15% of data are selected as training, validation, and test sets respectively.

4 Results

This paper presents results from both simulations based on the three models and practical implementation of RNN for flow rate measurement in an open channel flow loop.

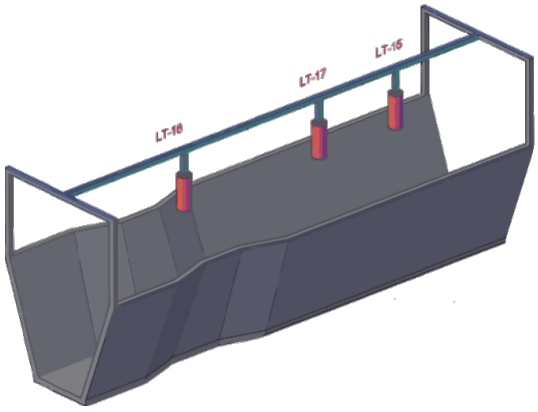
4.1 Simulation Study

RNN is implemented using all the three learning algorithms discussed in Section 2. Table 2 shows the optimal parameters used in the simulations. These optimal parameters are determined using grid search method and the optimization is done using Mean Absolute Percentage Error (MAPE). Apart from these parameters, number of neurons selected is 7, learning rate is 0.1 and number of folds for BPTT is 7.

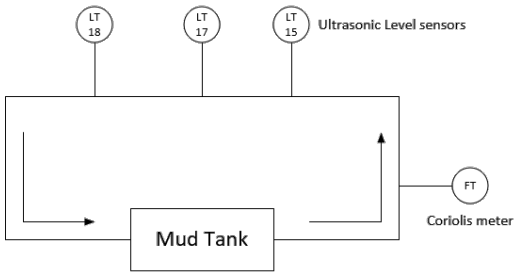
Figure 5 shows the comparison of RNN with different algorithms. As discussed in Section 2, EKF learning algorithm can quickly converge to a solution. From Figure 5 showing the MSE, it can be seen that EKF converges well before 20 epochs, BPTT converges around 100 epochs, and RTRL takes around 300 epochs to converge. The con-

Table 1. Input and output variables used for developing RNN models with the range and variable type.

Variables	*Range*	*Units*	*Type*
Upstream level measurement	31.2 - 107.5	mm	Input
Level measurement at the throat	28.9 - 78.3	mm	Input
Downstream level measurement	44.3 106.6	mm	Input
Mass flow rate	250 - 500	kg/min	Output



(a)



(b)

Figure 3. a) An open channel with Venturi section and three level sensors, LT-15, LT-17 and LT-18, with an arrow showing a flow direction. (b) Extremely simplified P&ID for the Venturi-rig flow loop with the measurands used in this study, viz. ultrasonic level sensors and FT-Coriolis mass flowmeter.

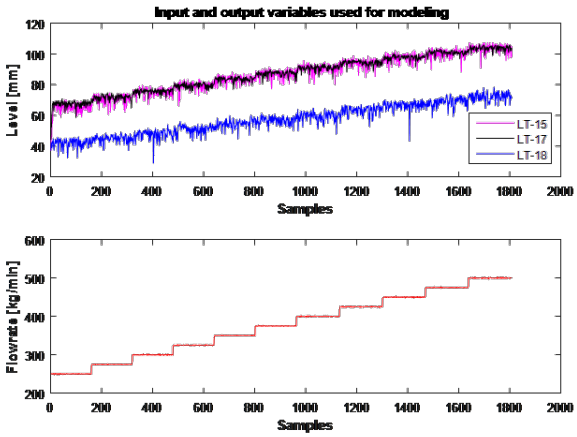


Figure 4. Input and output variables used for developing RNN models. First plot shows three level measurements with LT-15, LT-17 and LT-18. Second plot shows flowrate measurement using Coriolis mass flowmeter.

verging efficiency of these algorithms can be observed using the state parameters, which are weights of the neural network. Figure 6 shows the states of some of the weights while training a network. The state representation shows that the states in EKF and BPTT algorithms go to steady state very quickly. However, RTRL needs numerous training epochs for achieving steady states.

Figure 7 shows the estimations of different learning algorithms with reference to flow measurements from Coriolis mass flowmeter. The simulation results show that all the models using different learning algorithms are capable of describing the dynamics of the reference flow measurements well. RTRL has minimum MAPE out of the three models used, as shown in Table 2.

4.2 Experimental Study

The experimental study involves the implementation of simulation study in the Venturi rig. Despite slow convergence, RNN with RTRL learning algorithm is selected for its accuracy, less complexity, and faster computation. The algorithms for both BPTT and EKF have complex architectures and they are computationally demanding. This makes RTRL a suitable choice for implementing in the Venturi rig for the flow estimation. Figure 8 shows the experimental results obtained using model-drilling fluid in the test Venturi rig. The flow rate estimation using RNN is compared with the estimation previously made using static ANN and SVR. The comparison shows that RNN has better performance than other empirical models. The MAPE for RNN, ANN and SVR are 5.6%, 8.5%, and 7.7% respectively.

For the future work, we will try to improve the sensor measurements using suitable signal processing. As shown in Figure 4, the output mass flowrate using Coriolis mass flowmeter is less noisy as compared to the three input level measurements. Since the model completely depends on the data, we will work on online signal processing of level sensor measurements to reduce the noise in the measurements. In Figure 7 and Figure 8, we can see discontinuous peaks in the predictions of all the empirical models. By implementing these three models as an integral part of the processing algorithms (signal and control), we believe that

Table 2. Optimal parameters for different learning algorithms.

Learning algorithms	*Epochs*	*Number of previous inputs*	*Number of previous outputs*	*Mape [%]*
BPTT	200	1	3	2.97
RTRL	500	4	4	2.55
EKF	20	4	4	3.70

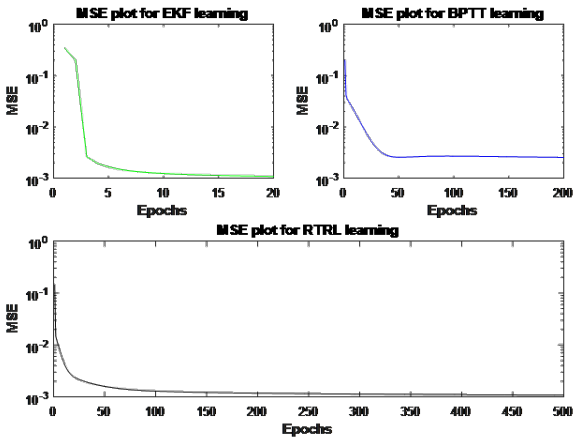


Figure 5. Mean Squared Error (MSE) plot for three different learning algorithms in RNN. Simulation results.

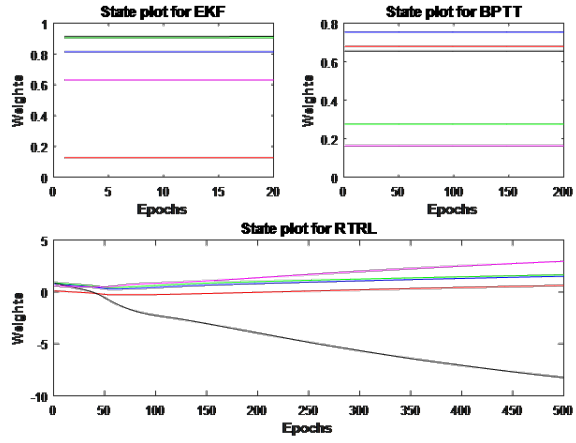


Figure 6. Different weights of the network in a state plot illustrating the convergence of the learning algorithms. Simulation results.

our model can be trained and operated with less noisy data resulting in improved predictions.

5 Conclusions

One way of having safe and efficient drilling operation is by continuously monitoring the properties of drilling mud. Any unwanted change in fluid properties can lead to two main problems; the influx of formation fluid and circulation fluid loss. The delta flow measurement while drilling is one of the best methods to detect the early influx or early fluid loss. In this paper, we introduced

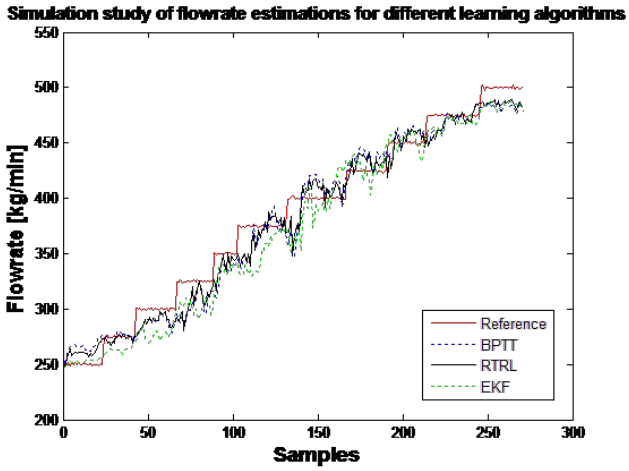


Figure 7. Comparison of flow rate estimation of Recurrent Neural Network (RNN) with three different learning algorithms with respect to Coriolis mass flow measurement as a reference measurement. Simulation results.

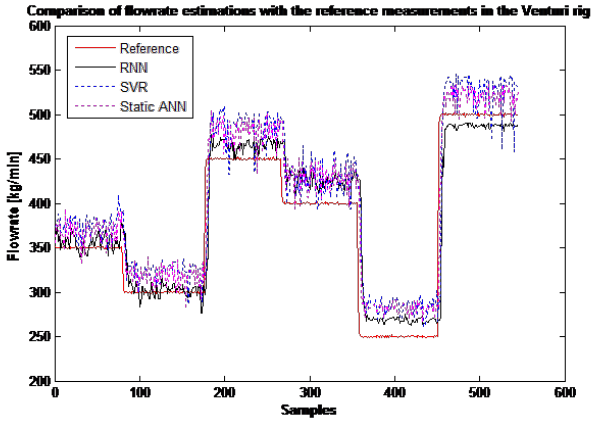


Figure 8. Comparison of flow rate estimations of a Dynamic Artificial Neural Network with Real Time Recurrent Learning algorithm (MAPE of 5.6%), a static Artificial Neural Network model (MAPE of 8.5%) and a Support Vector Regression model (MAPE of 7.7%) with respect to the Coriolis mass flow measurement as a reference flow measurement. Based on experiments using the Venturi-rig.

dynamic Artificial Neural Network to estimate the flow rate of non-Newtonian drilling fluids in an open channel venturi flume, which can be used for outflow measurement while determining delta flow. With Recurrent Neural Network, we simulated three different learning algorithms; Back Propagation Through Time, Real-Time Re-

current Learning and Extended Kalman Filter algorithm. The simulation results show that BPTT and EKF converge very quickly as compared to RTRL algorithms. Whereas, RTRL algorithm is more accurate, less complex and computationally faster than other two algorithms. So, based on this simulation analysis, RNN with RTRL algorithm is selected for the practical implementation. In the Venturi rig, RNN model with RTRL is implemented along with static ANN and Support Vector Regression (SVR) models. The experimental estimations of flow rates with respect to reference flow rate using Coriolis mass flowmeter show that the estimates based on RNN model has higher accuracy compared to ANN and SVR models. This improved performnce is due to the fact that RNN contains previous inputs and outputs as additional inputs for the current time, which are not considered in static ANN and SVR models. This study shows that non-intrusive ultasonic level measurements of the drilling fluid in an already existing open Venturi channel is a possible alternative to expensive devices such as Coriolis mass flowmeters to measure flow of drilling fluid.

Acknowledgement

The Ministry of Education and Research of the Norwegian Government is funding Khim Chhantyals PhD studies at University College of Southeast Norway (USN). Minh Hoang was partly involved in the development of this paper in conjunction with his master thesis at USN in 2016, (Hoang, 2016). The authors at USN appreciate the collaboration with and support from STATOIL for the assembly and commissioning of the open channel Venturi-rig with the various sensors and control system dedicated to the studies related to non-Newtonian fluids. We appreciate the expert advice on drilling operations by Dr. Geir Elseth of STATOIL. In addition, we acknowledge the practical work done by various groups of bachelor and master students of USN in conjunction with this project. Part of the work done is associated with the project SEMI-KIDD funded by the Research Council of Norway.

References

C. E. Agu and B. Lie. Numerical solution of the saint vernant equation for non-Newtonian fluid. In *Proceedings of the 55th Conference on Simulation and Modelling (SIMS 55), Modelling, Simulation and Optimization, 21-22 October 2014, Aalborg, Denmark*, number 108, pages 218–228. Linköping University Electronic Press, 2014a.

C. E. Agu and B. Lie. Smart sensors for measuring fluid flow using a venturi channel. In *Proceedings of the 55th Conference on Simulation and Modelling (SIMS 55), Modelling, Simulation and Optimization, 21-22 October 2014, Aalborg, Denmark*, number 108, pages 229–240. Linköping University Electronic Press, 2014b.

T. H. Ali, S. M. Haberer, I. P. Says, C. C. Ubaru, M. L. Laing, O. Helgesen, M. Liang, and B. Bjelland. Automated alarms for smart flowback fingerprinting and early kick detection. In *SPE/IADC Drilling Conference*. Society of Petroleum Engineers, 2013.

C. Berg, A. Tharanga, C. E. Agu, K. Chhantyal, and F. Mohammadi. *Simulatioin of open channel flow for mass flow measurement*. Technical report, University of South East Norway, Norway, 2013.

M. Boden. *A guide to Recurrent Neural Network and backpropagation*. Technical report, Halmstad University, Sweden, 2001.

D. B. Budik. *A Resource Efficient Localized Recurrent Neural Network Architecture and Learning Algorithm*. Technical report, University of Tennessee, USA, 2006.

R. Caenn, C. H. Darley, and G. R. Gray. *Composition and properties of drilling and completion fluids*. Gulf professional publishing, 2011.

K. Chhantyal, M. Hoang, H. Viumdal, and S. Mylvaganamand. Dynamic artificial neural network (dann) matlab toolbox for time series analysis and prediction. In *9th EUROSIM Congress on Modelling and Simulation, EUROSIM 2016*. SIMS Conference on Simulation and Modelling, 2016a.

K. Chhantyal, H. Viumdal, S. Mylvaganam, and G. Elseth. Ultrasonic level sensors for flowmetering of non-newtonian fluids in open venturi channels: Using data fusion based on artificial neural network and support vector machines. In *Sensors Applications Symposium (SAS), 2016 IEEE*, pages 1–6. IEEE, 2016b.

E. O. Dijk. *Analysis of Recurrent Neural Networks with application to speaker independent phoneme recognition*. Technical report, University Twente, Enschede, The Netherlands, 1999.

F. Frenzel. *Industrial flow measurement basics and practice*. Technical report, ABB automation products Gmbh, 2011.

G. Geratebau. *Equipment for engineering education, intruction manual HM 162.51 Venturi flume*. Technical report, Germany, 2013.

S. Hauge and K. Øien. Deepwater horizon: Lessons learned for the Norwegian petroleum industry with focus on technical aspects. *Chemical Engineering*, 26, 2012.

M. Hoang. *Viscosity measurement of non-Newtonian fluids*. Technical report, University of South East Norway, Norway, 2016.

M. Kamyab, S. R. Shadizadeh, H. Jazayeri-rad, and N. Dinarvand,. Early kick detection using real time data analysis with dynamic neural network: a case study in Iranian oil fields. In *Nigeria Annual International Conference and Exhibition*. Society of Petroleum Engineers, 2010.

P. Kim. *Kalman filter for beginners: with MATLAB examples*. CreateSpace, 2011.

M. W. Mak, K. W. Ku, and Y. L. Lu. On the improvement of the real time recurrent learning algorithm for recurrent neural networks. *Neurocomputing*, 24(1):13–36, 1999.

Appendix

List of symbols and abbreviations

Symbol	*Quantity*
ANN	Artificial Neural Network
BPTT	Back Propagation Through Time
CFD	Computational Fluid Dynamics
DANN	Dynamic Artificial Neural Network
EKF	Extended Kalman Filter
n	Number of folds
LT	Level Transmitter
MAPE	Mean Absolute Percentage Error
MSE	Mean Squared Error
N	Number of neurons
O	Order
P_b	Bottom hole pressure
P_f	Formation pressure
P_{ff}	Formation fracture pressure
RNN	Recurrent Neural Network
RTRL	Real Time Recurrent Learning
SVR	Support Vector Regression
t	Time

Dynamic Artificial Neural Network (DANN) MATLAB Toolbox for Time Series Analysis and Prediction

Khim Chhantyal Minh Hoang Håkon Viumdal Saba Mylvaganam

Faculty of Technology, Natural Sciences, and Maritime Sciences, University College of Southeast Norway,
{khim.chhantyal,hakon.viumdal,saba.mylvaganam}@usn.no, m.hoang1304@gmail.com

Abstract

MATLAB® Neural Network (NN) Toolbox can handle both static and dynamic neural networks. Using MATLAB® NN Toolbox with recurrent neural networks is not straight forward. We present a Dynamic Artificial Neural Network (DANN) MATLAB toolbox capable of handling fully connected neural networks for time-series analysis and predictions. Three different learning algorithms are incorporated in the MATLAB DANN toolbox: Back Propagation Through Time (BPTT) an offline learning algorithm and two online learning algorithms; Real Time Recurrent Learning (RTRL) and Extended Kalman Filter (EKF). In contrast to existing MATLAB® NN Toolbox, the presented MATLAB DANN toolbox has a possibility to perform the optimal tuning of network parameters using grid search method. Three different cases are used for testing three different learning algorithms. The simulation studies confirm that the developed MATLAB DANN toolbox can be easily used in time-series prediction applications successfully. Some of the essential features of the learning algorithms are seen in the graphical user interfaces discussed in the paper. In addition, installation guide for the MATLAB DANN toolbox is also given.

Keywords: dynamic artificial neural network (DANN), back propagation through time (BPTT), real-time recurrent learning (RTRL), extended Kalman filter (EKF), time series

1 Introduction

Artificial Neural Networks (ANN) are computational models consisting of many neurons in different layers with varying degrees of interconnections between them. The interconnection have weights assigned to them so that the ANNs can be tuned thus enabling them to learn and adapt. Feedforward or feedback networks are two broad classifications of ANNs. Feedforward ANNs use current inputs and current outputs, whereas, feedback ANNs use current and previous inputs and outputs. Feedback ANN performs time-series predictions and is a dynamic network. This type of network constitutes recurrent neural networks (RNN) either partially or fully connected depending on the extent of the feedback loops available in the network. Fully connected RNNs have interconnected feedback loops including self-feedback loops, whereas partially connected RNNs do not have self-feedback loops (Veelenturf and Gerez, 1999; Beale et al., 1992).

In an existing MATLAB® Neural Network Toolbox, there is a possibility to use feedforward ANN for static estimations and partially connected RNN for time-series predictions. This paper presents a MATLAB toolbox that can perform the empirical modeling using fully connected RNNs with three different learning algorithms. The following sections present the overview of the developed toolbox and the usage of the toolbox in three different practical applications.

2 Overview of Toolbox

The developed Dynamic Artificial Neural Network (DANN) toolbox consists of three main user interfaces, which are DANN Menu, Parameter Tuning, and Plot Menu. Each window consists of different elements as given below.

2.1 DANN Main

DANN Main is the main window of MATLAB DANN toolbox as shown in Figure 1. In this window, the user can upload the data set, divide data sets, select validation check, include bias, select learning algorithm, define learning parameters, select the number of previous inputs and outputs, and finally train the model.

2.1.1 Uploading data set

A user needs to upload his/her data set to train the model using MATLAB DANN toolbox. The format of the data set should be in '.mat' and each column should represent the variables in the model, where the last column is an output variable.

2.1.2 Division of data set

The uploaded data set should be divided into a training set, validation set, and testing set. A user can choose the percentage of data for training, validation and testing. Experience shows that 70% for the training set, 15% for both validation and testing works fine for any learning algorithms.

2.1.3 Validation check

A validation check is an option that prevents the over-fitting of the network. Over-fitting and under-fitting are most common problems encountered while dealing with

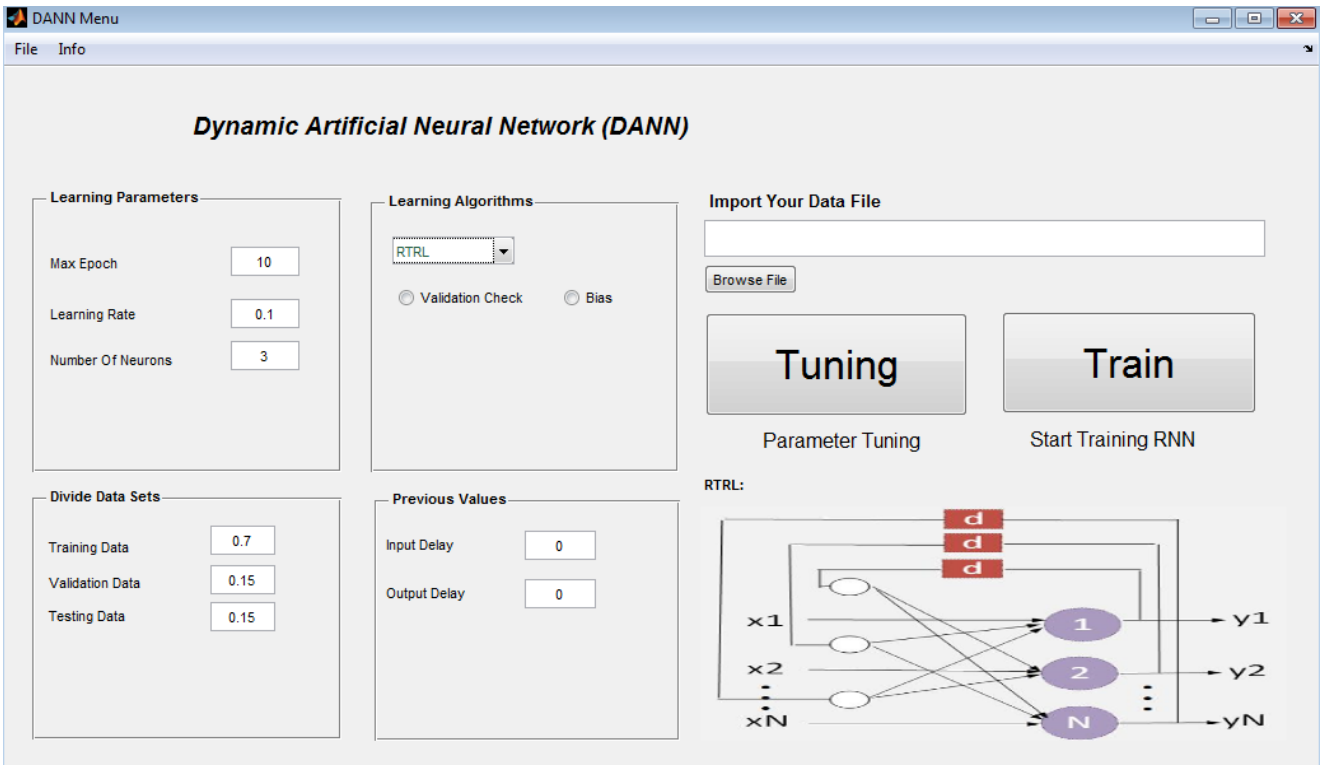


Figure 1. DANN Menu of MATLAB DANN toolbox: data upload, data division, selecting learning algorithms, and tuning learning parameters.

data models. Under-fitting can be improved by either tuning the learning rate or increasing the number of neurons in the network. The concept of over-fitting in MATLAB DANN toolbox is similar to the concept used in NN Toolbox in MATLAB® (Beale et al., 1992). The main idea is to terminate the RNN before the network gets over-fitted. For early stop, Mean Squared Error (MSE) for both training and validation is continuously monitored. While training a network, learning algorithm builds a certain hypothetical model for the network at each epoch.

The validation data are validated using the hypothetical model at that particular epoch. While learning, the value of MSE of training and validation data keep on reducing and the training of the network gets better with increasing epoch. However, the validation error can increase though the error for training decreases, which occurs in cases of over-fitting. The deterioration in the validation error can be attributed to the training process with random behavior (Beale et al., 1992). Therefore, the MATLAB DANN toolbox will count six consecutive increments in the validation error before it stops the learning algorithm.

When the model is over-fitted, the trained model seems to have good performance with training data, but it can have a large error while testing with the new data set (Beale et al., 1992). In other words, the trained model is not a generalized model when it is over-fitted. The implementation of validation check is presented in Case II in Section 3. In case, if the validation check is not selected, validation data will be part of the training data set.

2.1.4 Bias

It is an offset value added to the output of the neurons. It is often important to include bias in each neuron while constructing a neural network model. MATLAB DANN toolbox facilitates a choice to include or exclude bias terms in the network.

2.1.5 Learning algorithm

In this toolbox, there are three learning algorithms. The user can select any one of these algorithms based on the requirements regarding complexity, accuracy and application.

Back Propagation Through Time (BPTT) BPTT is an extension of gradient-based backpropagation algorithm that is used in feedforward ANN. The idea in BPTT is to unfold the RNN architecture into feedforward ANN architecture using an arbitrary number of folds. The BPTT architecture for the neural network with two neurons is shown in Figure 2. The network with BPTT algorithm is less complex compared to other learning algorithms. However, the complexity and the memory requirement increase when the number of folds increases. (Williams, 1992)

Real Time Recurrent Learning (RTRL) RTRL is one of most accepted real-time learning algorithms for RNN. In RTRL, the gradients at time 't' are computed using the propagation of gradients at previous time steps (Mak et al., 1999; Mandic and Chambers, 2000; Budik, 2006). The underlying RTRL architecture is shown in Figure 3,

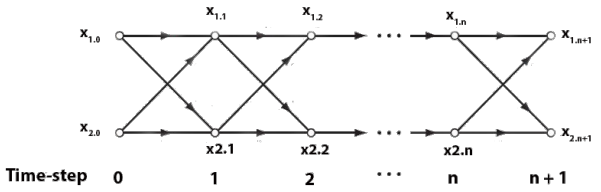


Figure 2. Architecture for Back Propagation Through Time (BPTT) learning algorithm with two neurons and n numbers of folding. (Haykin, 2009)

where **x**, **y**, N and d are inputs, outputs, number of neurons and unit time delay respectively. Based on the complexity, RTRL is the simplest online learning algorithm. However, the algorithm converges slowly and requires large memory for storage. (Williams, 1992)

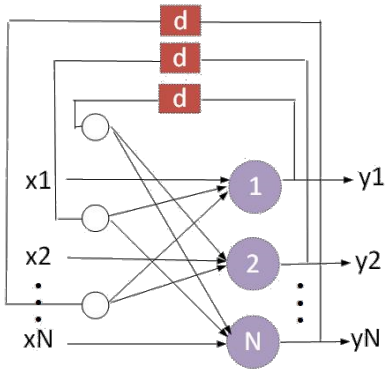


Figure 3. A general architecture for Real Time Recurrent Learning (RTRL) and Extended Kalman Filter (EKF) learning algorithms showing self-feedback and feedback loops within the neurons. Vectors **x** and **y** as input and output with d as the delay.

Extended Kalman Filter Learning (EKF) EKF can be used as a supervised on-line learning algorithm to tune the weights of RNN. In EKF, the state vector consists of weights and the locally induced outputs of each neuron in the network. Regarding convergence, EKF is the fastest algorithm among the algorithms presented in the MATLAB DANN toolbox. The order of computational complexity for EKF is the same as for RTRL, and the storage requirement is larger for EKF. The general architecture for EKF learning is shown in Figure 3. (Williams, 1992)

The main problem using gradient-based learning algorithms is vanishing gradient problem. As a solution to this problem, the German researcher Sepp Hochreiter and Juergen Schmidhuber introduced recurrent net with Long Short-Term Memory (LSTM) units (Haykin, 2009). In recent publications (Sak et al., 2014; Zen and Sak, 2015), LSTM RNN architectures are implemented because of their accuracy.

2.1.6 Learning parameters

The parameters of learning algorithms such as the number of neurons, learning rate, the maximum number of epochs and number of folds are discussed in this section.

Number of neurons The neurons and the connections between the neurons are essential features of a neural network. The number of neurons plays a vital role in the performance of the neural network. Too few neurons may not completely describe the dynamics of the system, and too many neurons can increase the complexity of the network (Haykin, 2009; Siddique and Adeli, 2013). Therefore, an optimal selection of a number of a neuron is one of the most important aspects of neural network modeling. In MATLAB DANN toolbox, each neuron is associated with the sigmoid function with the range [0, 1].

Learning rate The learning rate determines the rate of learning of gradient-based learning algorithms like BPTT and RTRL. The range of learning rate is [0, 1] and determines the converging efficiency while learning. The very small value of learning rate will slow down the learning algorithm and may require a large number of epochs to converge to a solution. Whereas the high value of learning rate can converge quickly, but it might have large variations and fluctuations in MSE of a training data. (Haykin, 2009; Siddique and Adeli, 2013)

Maximum number of epoch In MATLAB DANN toolbox, there are two stopping criterions. One of them is validation check, which is already discussed. Another way of stopping the training is the maximum number of epochs. A user can select a maximum number of epochs for the training using DANN Menu.

Number of folds The number of folds is a parameter for BPTT learning. The default selection is '3', which is the minimum possible value that can be selected for a given number of folds.

2.1.7 Past inputs and outputs

In applications involving prediction of time-series, the current output depends on the previous inputs and outputs. MATLAB DANN toolbox allows a user to select a number of previous inputs and previous outputs as additional inputs to find the output at the current time. By default, if the values are selected as '0' for both input and output, MATLAB DANN toolbox will use one previous output from each neuron.

2.1.8 Additional parameters

In EKF learning algorithm, a user must assign three more parameters for learning, which are Sigma_U, Sigma_W, and Sigma_O in MATLAB DANN toolbox. These parameters are tuning parameters for the output of each neuron as a state, weights as a state and output of the network respectively. These parameters are responsible for Kalman gain calculation for the states (i.e. output of neuron and weight) and determine the update of the output of each neuron and the weight connections between the neurons (Williams, 1992). As the simulation stops, the parameters, weights and other information regarding the simulation are saved in the workspace in MATLAB®.

2.2 Parameter Tuning

In any implementation of ANN, tuning of parameters is one of the biggest challenges. The optimal selection of network parameters can only lead to a good model. Contrary to existing MATLAB® Neural Network Toolbox, MATLAB DANN toolbox has a facility to tune the parameters optimally. In DANN Main, if you click on Tuning button, Parameter Tuning window will open as shown in Figure 4. The optimal tuning is based on the grid search method, and optimality is evaluated using Mean Absolute Percentage Error (MAPE). In the left panel of the window, a user can assign lower limit, higher limit and an increment to each parameter and start tuning. At the end of the tuning, optimal values of the parameters are displayed in the right panel of the window with minimum MAPE. Usually, parameter tuning takes a long time, so MATLAB DANN toolbox provides an option to get notification alarm. It is to be noted that a user must upload data, select learning algorithm, decide to or not to include bias and validation check before starting the tuning process. Thus, obtained optimal parameters can be used for training the model.

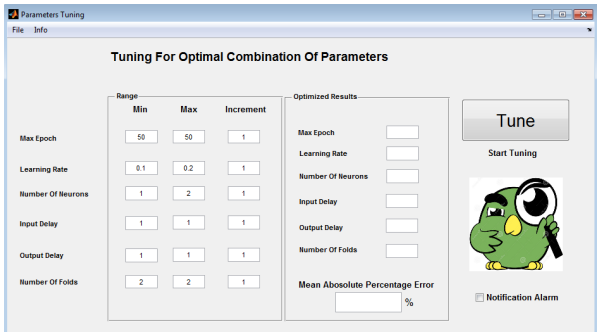


Figure 4. Parameter Tuning window of MATLAB DANN toolbox that allows a user to tune the optimal parameters based on grid search method.

2.3 Plot Menu

Plot Menu window pops-up when the simulation is completed as shown in Figure 5. It consists of five different types of plots, which are performance plot, regression plot, prediction plot, parameter plot, and error plot.

2.3.1 Performance plot

It shows the MSE for training data set and validation error for each epoch.

2.3.2 Regression plot

It compares the target output and model prediction in terms of squared correlation coefficient such that '0' meaning not related at all and '1' meaning highly correlated to each other.

2.3.3 Prediction plot

It shows the test data and model prediction with MAPE between them.

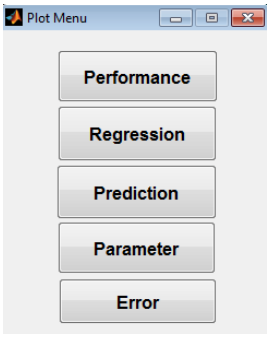


Figure 5. Plot Menu of MATLAB DANN toolbox with different plots for the analysis of the model.

2.3.4 Parameter plot

It shows the states of five different randomly chosen weights at different epochs. The analysis using parameter plot is very efficient if you are working with some system identification problems. In that case, one can visualize how the weights change with epochs. The steady state values of the weights after some epochs are the model parameters in typical system identification problems.

2.3.5 Error plot

It shows the error between the target value and the model prediction for each test samples.

2.4 Additional information

The MATLAB DANN toolbox has additional help options for the users. A user can get general information in Q&A section inside the Help Window. With a right-click in any parameter name, action buttons or selection options, a prompt help window related to that expression will pop up.

The MATLAB DANN toolbox in the current version has the following limitations:

(a) The toolbox is not yet ready for linking to Simulink.

(b) The toolbox handles Multiple Input Single Output (MISO) scenario and needs some modification to address Multiple Input Multiple Output (MIMO) scenarios.

(c) The execution time needed for the current version can be reduced.

2.5 Installing the MATLAB DANN toolbox

The first step in installing MATLAB DANN toolbox is to download installation file. Double-click in the downloaded installation file will direct to the installation process in the MATLAB®. It is recommended to have a MATLAB® version 2014 or later.

3 Case Studies

In this section, the usage of the MATLAB DANN toolbox in three different practical applications is discussed.

These three different cases use the data set from an experimental flow rig, and example data sets from MATLAB® Neural Network Toolbox. To give a better understanding in analyzing the simulation results, different sets of plots are investigated under these cases.

3.1 Case I: BPTT learning algorithm for flow measurement

In drilling operations, the flow rates of drilling mud at inflow and outflow positions can be used to detect kick and fluid loss. An open channel flow loop is available at University College of Southeast Norway (USN) for the study of outflow measurement. The data set with three level measurements as inputs and a flow measurement as the single output are taken from the flow loop for the analysis of BPTT learning algorithm in MATLAB DANN toolbox. Figure 6 and Figure 7 show the regression plot and prediction plot for flow estimation using BPTT learning algorithm in the toolbox. The simulation results show that the BPTT learning algorithm provided by MATLAB DANN toolbox is capable of mapping the inputs and outputs with high accuracy.

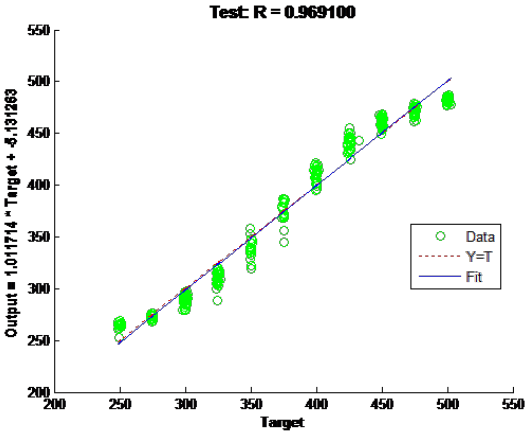


Figure 6. The regression plot for flow measurement using BPTT learning algorithm, with a correlation of 96% between the target values and the model prediction values. Data set from an experimental flow rig at USN.

3.2 Case II: EKF learning algorithm for temperature measurement

To analyze the performance of EKF learning algorithm, an example data set provided by MATLAB® Neural Network Toolbox is used. The data set of a liquid-saturated steam heat exchanger consists of time-series liquid flow rate and liquid outlet temperature, used as input and output to the ANN feedback network respectively. Figure 8 and Figure 9 show the performance plot and prediction plot for fully connected RNN with EKF learning algorithm. The learning algorithm has an early stop at 8 epochs due to the validation check with MSE of 0.015. The low value of MAPE in prediction plot shows that the EKF learning algorithm with validation check is able to generalize the model and avoid over-fitting.

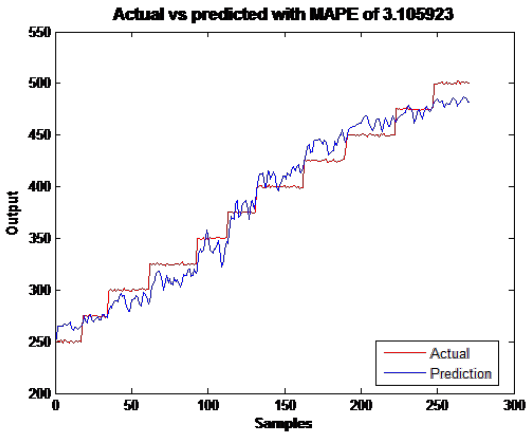


Figure 7. The prediction plot for flow measurement using BPTT learning algorithm with a MAPE of 3.1%. Data set from an experimental flow rig at USN.

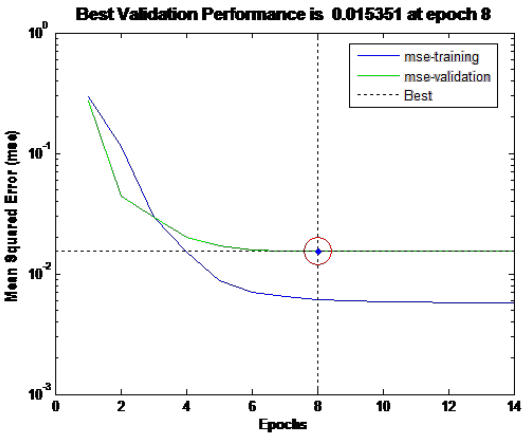


Figure 8. The performance plot for temperature measurement using EKF learning algorithm. The best validation performance is 0.015351 at epoch 8. Data set from MATLAB® Neural Network Toolbox.

3.3 Case III: RTRL learning algorithm for mortality prediction

Another example data set provided by MATLAB® Neural Network Toolbox is used to investigate the performance of RTRL learning algorithm. The data set is a Pollution mortality data set that consists of eight input variables (Temperature, Relative humidity, Carbon monoxide, Sulfur dioxide, Nitrogen dioxide, Hydrocarbons, Ozone, and Particulates) and total mortality as an output variable. Figure 10 to Figure 13 shows the simulation results for mortality prediction using RTRL learning algorithm using MATLAB DANN toolbox.

The parameter plot as shown in Figure 10 shows the states of randomly chosen weights of the network. As discussed in Section 2, it takes longer time for the weights to converge to a steady state when using RTRL.

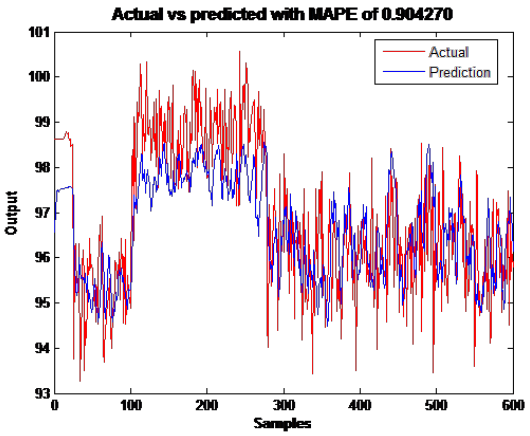


Figure 9. The prediction plot for temperature measurement using EKF learning algorithm with a MAPE of 0.9%. Data set from MATLAB® Neural Network Toolbox.

The regression plot as in Figure 11 illustrates that the predictions using RTRL are highly correlated with the target values with a correlation of 92%. The MAPE between the predicted values and target values are 2.93% as shown in Figure 12. The error in each sample is shown in the error plot in Figure 13.

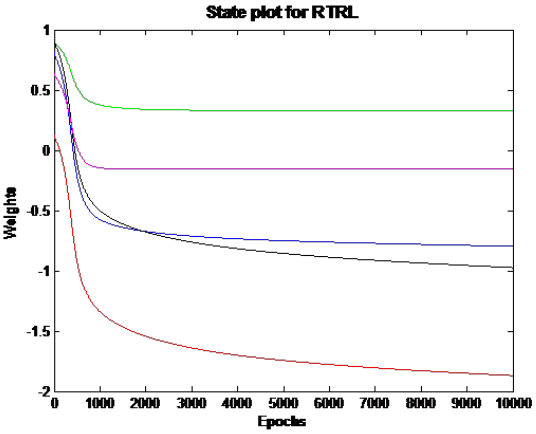


Figure 10. The state plot for mortality time-series prediction using RTRL learning algorithm. Data set from MATLAB® Neural Network Toolbox.

4 Conclusions

The existing MATLAB® Neural Network Toolbox has a possibility to use both static and dynamic neural networks. However, it is not possible directly use the toolbox to fully connected recurrent neural networks. For this reason, this study presents the Dynamic Artificial Neural Network MATLAB toolbox that gives an opportunity to use the fully connected neural network for time-series predictions. The toolbox consists of three different learning algorithms, where Back Propagation Through Time (BPTT) is an offline learning algorithm, Real Time Recurrent Learning (RTRL) and Extended Kalman Filter (EKF)

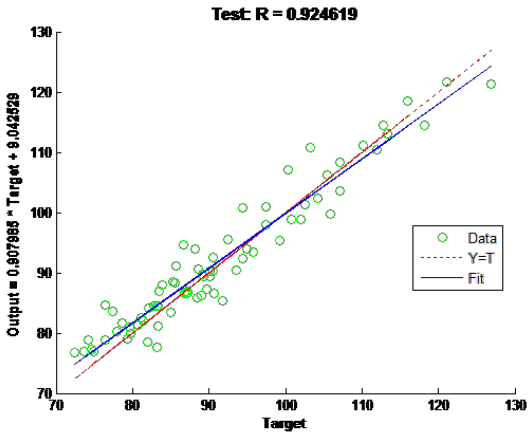


Figure 11. The regression plot for mortality time-series prediction using RTRL learning algorithm with 92% correlation between target values and model predictions. Data set from MATLAB® Neural Network Toolbox.

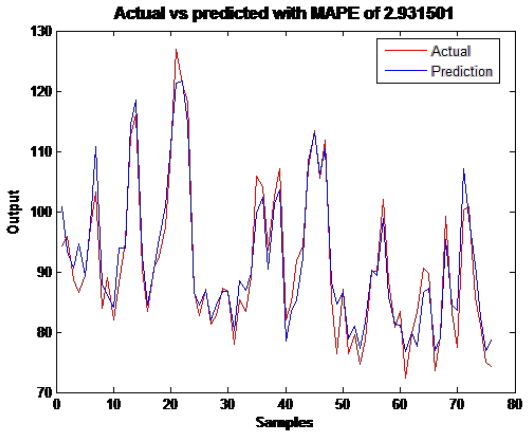


Figure 12. The prediction plot for mortality time-series prediction using RTRL learning algorithm with MAPE of 2.93%. Data set from MATLAB® Neural Network Toolbox.

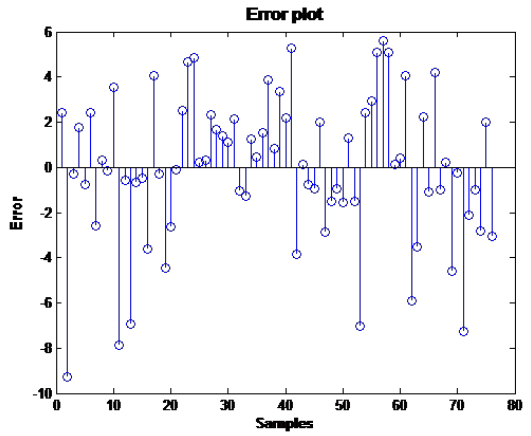


Figure 13. The error plot for mortality time-series prediction using RTRL learning algorithm with 9 units as the highest error in the test samples. Data set from MATLAB® Neural Network Toolbox.

learning algorithm are online learning algorithms. Main details and guides for installing and using the developed toolbox are presented in this paper.

To demonstrate the features of the MATLAB DANN toolbox, three different practical problems are considered using three different learning algorithms. The simulation studies presented in this paper show that the developed toolbox can be used in applications involving time-series predictions. In addition, the developed toolbox has a dedicated option for the optimal tuning of parameters.

This Toolbox can be used in financial market trending studies with some modification similar to (Pelusi et al., 2014).

This work is meant for academic use with particular focus on the students using the existing MATLAB® Neural Network Toolbox. In future, other different learning algorithms can be included in the developed toolbox with some programming efforts.

Acknowledgement

The Ministry of Education and Research of the Norwegian Government is funding Khim Chhantyal's PhD studies at University College of Southeast Norway (USN). Minh Hoang was partly involved in the development of this paper in conjunction with his master thesis at USN in 2016, (Hoang, 2016). The authors at USN appreciate the collaboration with and support from STATOIL for the rig used in the current studies for generation of time series of flow rates. We appreciate the expert advice on drilling operations by Dr. Geir Elseth of STATOIL. In addition, we acknowledge the practical work done by various groups of bachelor and master students of USN in conjunction with this project. Part of the work done is associated with the project SEMI-KIDD funded by the Research Council of Norway.

References

Daniel Borisovich Budik. *A Resource Efficient Localized Recurrent Neural Network Architecture and Learning Algorithm*. Master Thesis, University of Tennessee, 2006.

Danilo P Mandic and Jonathon A Chambers. A normalised real time recurrent learning algorithm. *Signal processing*, 80(9):1909–1916, 2000. doi:10.1016/S0165-1684(00)00101-8.

Danilo Pelusi, Massimo Tivegna, and Pierluigi Ippoliti. Intelligent algorithms for trading the euro-dollar in the foreign exchange market. In *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pages 243–252. Springer, 2014. doi:10.1007/978-3-319-02499-822.

Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Heiga Zen and Hasim Sak. Undirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pages 4470-4474. IEEE, 2015. doi: 10.1109/ICASSP.2015.7178816.

r LPJ Veelenturf and Ir SH Gerez. Analysis of recurrent neural networks with application to speaker independent phoneme recognition. 1999.

Man-Wai Mak, Kim-Wing Ku, and Yee-Ling Lu. On the improvement of the real time recurrent learning algorithm for recurrent neural networks. *Neurocomputing*, 24(1):13–36, 1999. doi: 10.1016/S0925-2312(98)00089-7.

Mark Hudson Beale, Martin T Hagan, and Howard B Demuth. *Neural Network Toolbox^{TM} User's Guide*. 1992.

Minh Hoang. *Viscosity measurement of non-Newtonian fluids*. Master Thesis, University of South East Norway, Norway, 2016.

Nazmul Siddique and Hojjat Adeli. *Computational intelligence: synergies of fuzzy logic, neural networks and evolutionary computing*. John Wiley & Sons, 2013.

Ronald J Williams. *Some observations on the use of the extended Kalman filter as a recurrent network learning algorithm*. College of Computer Science, Northeastern University, 1992.

Simon Haykin. *Neural networks and learning machines*, volume 3. Pearson Upper Saddle River, NJ, USA, 2009.

Simulation of Bubbling Fluidized Bed using a One-Dimensional Model Based on the Euler-Euler Method

Cornelius Agu* Marianne Eikeland Lars Tokheim Britt Moldestad

Department of Process, Energy and Environmental Technology
University College of Southeast Norway, Norway, {Cornelius.e.agu, Marianne.Eikeland, Lars.A.Tokheim, britt.moldestad}@usn.no

Abstract

The behaviour of a fluidized bed can be modeled based on the Euler-Euler approach. This method has been fully utilized in both three-dimensional (3D) and two-dimensional (2D) systems for obtaining, for example, the axial and radial distribution of fluidized bed properties. However, the bed property such as void fraction distribution along the flow direction can be of great interest for a design purpose. To save computational cost, an appropriate one-dimensional (1D) model can be used to obtain the average bed property along the vertical axis of a fluidized bed. In this paper, a 1D model based on the Euler-Euler method is presented. The results show that the model can be used to describe the behaviour of a fluidized bed. With a reasonable accuracy, the results also show that the 1D model can predict the minimum fluidization velocity and the superficial gas velocity at the onset of slugging regime.

Keywords: Euler-Euler, bubbling, void fraction, fluidized bed, flow regime

1 Introduction

The fluidized bed has wide industrial applications. Such applications include circulation of catalyst particles in a chemical reactor, pneumatic transport of particles and gasification of coal/biomass. In fluidized bed reactors, there is a good mixing of solids and fluid, and this enhances heat and mass transfer rates between the fluid and the particles.

For the purpose of design and prediction of hydrodynamic behaviour of fluid-particle systems, several empirical and semi-empirical models have been developed. Moreover, the computational fluid dynamics has also been applied in such a multi-phase system. As in a single-phase system, the mass, momentum and energy transfers also govern the motions of fluid and particles in the bed. The interface momentum transfer between the phases influences the behaviour of the system. When a fluid flows through a bed of particles, the drag force acts continuously against the weight of the bed. At a certain fluid velocity, the bed begins to float in the fluid stream. This velocity is generalized as the minimum fluidization velocity. Previous studies have shown that at this fluid velocity, the interphase drag force corresponds to the net weight of the bed. This concept is used in deriving models for estimating the minimum fluidization velocity from the drag models (Kunnii and Levenspiel, 1991). Due to complexities arising from particle-particle interactions and particle-wall interactions, it has been proven difficult to establish accurate fluid-particle interphase drag models to predict accurately the behaviour of fluidized beds. However, a number of drag models can be found in the literature (Taghipour et al, 2005; Beuzarti and Bournot, 2012; Li et al, 2009).

Beyond the onset of bed fluidization, and with increasing superficial gas velocity, the agitation of particles in the bed increases. Different phase transitions can be observed when a bed is fluidized. As the fluid velocity increases, a fluidized bed passes through the bubbling regime, the turbulent regime, fast fluidization and the pneumatic conveying regime (Kunnii and Levenspiel, 1991).

In this study, the focus is on modelling a bubbling fluidized bed. A number of models have been developed for such a regime. Davidson and Harrison (1965) developed a simple two-phase model based on a mass balance and experimental observations. The underlying assumption in this model is that two distinct phases, – bubble and emulsion exist throughout the bed. A more advanced model based on physics of mass, momentum and energy conservations have also been developed. Two widely used approaches to this model development are those based on the Euler-Euler and the Euler-Lagrange methods (Crowe et al, 2012). Depending on the fluid-particle drag model and the numerical method employed, the two- and three-dimensional (2D and 3D) versions of these models have been proven successful in predicting the behaviour of fluid-particle multiphase systems. One major drawback is that the 2D and 3D models are highly computational time demanding.

There is a limited number of studies based on a 1D model. Solsvik et al (2015) used a 1D model in a methane reforming studies, and Silva (2012) presented a non-conservative version of the model for simulating

the bubbling bed behaviour of a biomass gasification process.

In this paper, the goal is to develop a detailed one-dimensional model that predicts well the behaviour of a fluidized bed with less computational time. A 1D model based on the Euler-Euler approach is used to study the behaviour of glass bead particles in a bubbling bed. The simulated results are compared with experimental data obtained from a cold fluidized bed, and with the simulation results based on a three dimensional model. The simulated superficial gas velocity at the onset of slugging is compared with the result obtained from the correlation (Geldart, 1986).

2 Computational Model

2.1 Governing Equations

The governing equations for the motions of fluid and particles in a fluidized bed are developed based on the Euler approach, and are given in (1) – (5). In the following, the subscripts "s" and "g" denote solid and gas. u and v are the respective gas and particle velocities, g is the acceleration due to gravity, β_d is the momentum transfer coefficient, and P, ε and ρ are the pressure, volume fraction and density, respectively. f is the wall frictional factor.

2.1.1 Continuity Equations

$$\frac{\partial}{\partial t}\left(\varepsilon_g\rho_g\right) + \frac{\partial}{\partial z}\left(\varepsilon_g\rho_g u\right) = 0 \tag{1}$$

$$\frac{\partial}{\partial t}\left(\varepsilon_s\rho_s\right) + \frac{\partial}{\partial z}\left(\varepsilon_s\rho_s v\right) = 0 \tag{2}$$

$$\varepsilon_g + \varepsilon_s = 1 \tag{3}$$

2.1.2 Momentum Equations

$$\frac{\partial}{\partial t}\left(\varepsilon_g\rho_g u\right) + \frac{\partial}{\partial z}\left(\varepsilon_g\rho_g u.u\right) = \frac{\partial}{\partial z}\left(\mu_{eg}\frac{\partial u}{\partial z}\right) - \varepsilon_g\frac{\partial P_g}{\partial z} - \frac{2f_g\varepsilon_g\rho_g u|u|}{D_h} - \varepsilon_g\rho_g g + \beta_d(v - u) \tag{4}$$

$$\frac{\partial}{\partial t}\left(\varepsilon_s\rho_s v\right) + \frac{\partial}{\partial z}\left(\varepsilon_s\rho_s v.v\right) = \frac{\partial}{\partial z}\left(\mu_{es}\frac{\partial v}{\partial z}\right) - \varepsilon_s\frac{\partial P_g}{\partial z} - \frac{2f_s\varepsilon_s\rho_s v|v|}{D_h} - \varepsilon_s\rho_s g - \frac{\partial P_s}{\partial z} + \beta_d(u - v) \tag{5}$$

Here, $D_h = 4A/P_{wet}$ is the bed hydraulic diameter, where A is the bed cross-sectional area and P_{wet} is the wetted perimeter of the bed. $\mu_{es} = 2\mu - \lambda$ is the phase equivalent dynamic viscosity. The solid pressure and solid stress due to collisions are based on the kinetic theory of granular flow. The constitutive equations of the model (1) – (5) are given in (6) - (10).

2.1.3 Constitutive Equations

- Gas phase (Gidaspow, 1994)

$$f_g = \begin{cases} 16Re_g^{-1} ; & Re_g \le 2300 \\ 0.0791Re_g^{-0.25} ; & Re_g > 2300 \end{cases} \tag{6}$$
$$Re_g = \varepsilon_g\rho_g uD/\mu_g$$

- Solid phase (Gidaspow, 1994; Lathowers and Bellan, 2000)

$$f_s = 0.048|v|^{-1.22} \tag{7}$$

$$P_s = K_1\varepsilon_s^2\theta \tag{8}$$

$$\lambda_s = K_2\varepsilon_s\sqrt{\theta} \tag{9}$$

$$\mu_s = K_3\varepsilon_s\sqrt{\theta} \tag{10}$$

where

$K_1 = 2(1 + e)\rho_s g_0$,

$K_2 = \frac{4d_p\varepsilon_s\rho_s g_0(1 + e)}{3\sqrt{\pi}} - \frac{2}{3}K_3$,

$K_3 = d_p\rho_s/2\left[\frac{\sqrt{\pi}}{3(3 - e)}\left\{1 + 0.4\varepsilon_s g_0(1 + e)(3e - 1)\right\} + \frac{8\varepsilon_s g_0(1 + e)}{5\sqrt{\pi}}\right]$,

$g_0 = \frac{3}{2}\left[1 - \left(\frac{\varepsilon_s}{\varepsilon_{maxP}}\right)^{1/3}\right]^{-1}$,

$\theta = \left[\frac{\left(\sqrt{(K_1\varepsilon_s)^2 + 4K_4\varepsilon_s(K_2 + 2K_3)} - K_1\varepsilon_s\right)}{2K_4\varepsilon_s}\frac{\partial v}{\partial z}\right]^2$,

$K_4 = 12\rho_s g_0(1 - e^2)/(d_p\sqrt{\pi})$.

Here, θ is the granular temperature, g_0 is the radial distribution function, e is the coefficient of restitution and d_p is the single particle diameter. ε_{maxP} is the solid fraction at maximum packing with a value of about 0.7406. μ and λ are shear and bulk viscosity, respectively.

2.2 Drag Model

There are number of drag models that can be found in literature. In this paper, the model proposed by Gidaspow (1994) is used.

$$\beta_d = \begin{cases} \beta_{dErg} ; & \varepsilon_g \le 0.8 \\ \beta_{dWY} ; & \varepsilon_g > 0.8 \end{cases} \tag{11}$$

Here, β_{dErg} and β_{dErg} are given by (12) and (13), respectively.

$$\beta_{dErg} = 150\frac{\varepsilon_s^2\mu_g}{\varepsilon_g(\phi_s d_p)^2} + 1.75\frac{\varepsilon_s\rho_g|u - v|}{\phi_s d_p} \tag{12}$$

$$\beta_{dWY} = \frac{3}{4}C_d\frac{\varepsilon_s\varepsilon_g\rho_g}{\phi_s d_p}|u - v|\varepsilon_g^{-2.65} \tag{13}$$

where

$$C_d = \begin{cases} \frac{24}{Re_p}\left(1 + 0.15Re_p^{0.687}\right); & Re_p<1000 \\ 0.44; & Re_p\geq 1000 \end{cases},$$

$$Re_p = \frac{\varepsilon_g \rho_g |u - v|}{\mu_g} d_p.$$

C_d is the drag coefficient and Re_p is the particle Reynolds number. $\emptyset_s$ is the single particle sphericity. To avoid discontinuity in using the above drag model, a weighting function proposed by Lathowers and Bellan (2000) is used.

$$\beta_d = (1 - \omega_d)\beta_{dErg} + \omega_d \beta_{dWY} \quad (14)$$

$$\omega_d = \frac{1}{\pi}\tan^{-1}\left(150*1.75\left(0.2 - \left(1 - \varepsilon_g\right)\right)\right) + 0.5 \quad (15)$$

2.3 Void Fraction Equation

Another crucial issue is the prediction of void fraction ε_g along the bed. It is obvious that neither (1) nor (2) can predict the void if used alone. This is due to the dependency of void fraction on the relative velocity between the solid particles and the fluid. In the computer code MFIX, the solid volume fraction is obtained based on a guess-and-correction method (Syamlal, 1998). Effective application of this method requires a known function of solid pressure with the solid volume fraction.

With the assumption that both solid particles and fluid have a constant density over the bed, the void fraction equation is established based on (1) and (2) (Gidaspow, 1994). However, due to changes of fluid pressure in the bed, there could be slight changes in the fluid density, which may influence the bed behaviour. In this paper, a new version of the void equation developed based on the continuity equations for gas and solid phases, is introduced. The new void equation, described below, partially accounts for the effect of fluid density variation.

$$\alpha_v \frac{\partial \varepsilon_g}{\partial t} + v_m \frac{\partial \varepsilon_g}{\partial z} = \varepsilon_s \varepsilon_g \rho_{rg} \frac{\partial v_r}{\partial z} \quad (16)$$

Here, $v_r = v - u$ is the relative velocity between the solid particles and the fluid. v_m and α_v are mixture mass velocity and relative volume fraction, respectively, and are expressed as

$$\alpha_v = \varepsilon_g \rho_{rg} + \varepsilon_s, \quad (17)$$

$$v_m = \varepsilon_g \rho_{rg} v + \varepsilon_s u. \quad (18)$$

where, $\rho_{rg} = \rho_g/\rho_{ref}$ is the reduced gas density. The gas density is obtained, assuming the ideal gas behaviour, $\rho_g = \frac{P_g}{RT}$.

2.4 Minimum Fluidization Velocity

The onset of fluidization occurs at a certain velocity where the net weight of the bed balances the drag force between the fluid and the bulk of particles in the bed. The minimum fluidization velocity, U_{mf} can be obtained from

$$U_{mf} = \frac{\mu_g}{\rho_g d_p} Re_{p.mf}. \quad (19)$$

The particle Reynolds number at minimum fluidization condition $Re_{p.mf}$ is based on the Ergun's bed pressure drop model (Ergun, 1952),

$$150\frac{(1 - \varepsilon_{mf})}{\varepsilon_{mf}^3 \emptyset_s^2} Re_{p.mf} + 1.75\frac{1}{\varepsilon_{mf}^3 \emptyset_s} Re_{p.mf}^2 = Ar, \quad (20)$$

where Ar is the Archimedes number, expressed as

$$Ar = \frac{d_p^3 \rho_g (\rho_s - \rho_g) g}{\mu_g^2}. \quad (21)$$

Here, ε_{mf} is the bed void fraction at the minimum fluidization condition. U_{mf} and ε_{mf} are bed properties, and either of them must be known for the other to be calculated from (19) – (21). A number of empirical correlations for ε_{mf} are available (Kunnii and Levenspiel, 1991), but this paper uses the correlation proposed by Wen and Yu (1966).

$$\frac{1}{\emptyset_s \varepsilon_{mf}^3} \approx 14 \quad (22)$$

3 Experimental setup

The experimental setup consists of a vertical cylindrical column of height 1.4 m and base diameter 0.084 m. The rig is fitted with ten pressure sensors, measuring the fluid pressure in the column up to the height of about 1.0 m. Compressed air at ambient temperature is used as the fluidizing medium. The bottom of the column is fitted with a porous plate. The porous plate ensures even distribution of air within the bed.

Thapa and Halvorsen (2013) conducted experiments with this cold fluidized bed rig using glass beads particles (particle size 350 μm) at a bed height of 0.32 m (see Figure 1). The experimental data used in this paper are those reported in Thapa and Halvorsen (2013).

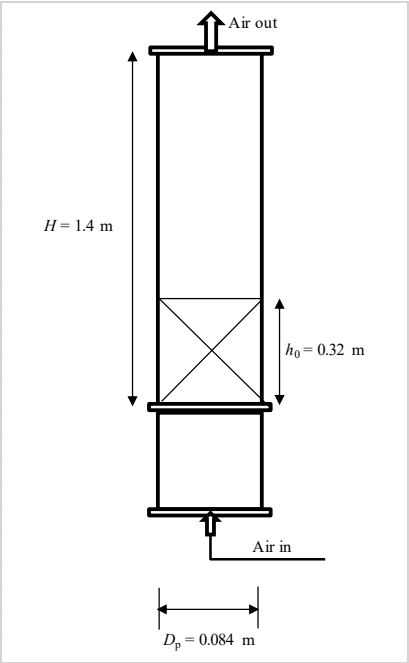


Figure 1. Physical Dimension of the fluidized bed column.

4 Simulations

The solution of the model described in Section 2 for the fluid-particle system is based on the finite volume method with staggered grids. The models are discretized in space using the first order upwind scheme, and in time based on the implicit method. The SIMPLE algorithm is used for the pressure-velocity coupling. The entire codes for the system are implemented and run in MATLAB. The properties of fluid and particles used in the computation are summarized in Table 1.

4.1 Fluidized bed regimes

In addition to simulating a bubbling fluidized bed, the transitions between different regimes for a fluidized bed are simulated using the 1D model. The flow transition from one regime to another depends on a number of factors. These include the bed particle size, the size distribution, the superficial gas velocity and the relative size between the bed height and the bed diameter. For a bed with Geldart B particles, the particle size and size distribution do not influence slugging in the bed (Baeyens and Geldart, 1974). As given in Yang (2003), slugging will occur if $\frac{h_0}{D_h} > 2$. The minimum gas velocity for the onset of slugging can be obtained from (23) (Geldart, 1986) as used in Xie et al (2008).

$$U_{ms} = U_{mf} + 0.0016(60D_t^{0.175} - h_{mf})^2 + 0.07(gD_t)^{0.5} \quad (23)$$

Here, all the length units are expressed in (cm), and h_{mf} is the bed height at minimum fluidization condition.

4.2 Initial and Boundary Conditions

Initially, the fluid pressure distribution is assumed hydrostatic, and the fluid velocity is considered uniform throughout the column, as described in Table 2. The inlet fluid pressure is assumed fixed, and it corresponds to the total weight of particles in the bed. Since the focus is on a bubbling bed, the outlet solid volume fraction is fixed to zero, while the fluid pressure at exit is taken to be atmospheric. The inlet boundary value for the solid volume fraction is dynamic, and then obtained appropriately from the void propagation equation.

5 Results and Discussion

Thapa and Halvorsen (2013) used the experimental rig described above to study the fluid-particle behaviour in a bed with particles having an average diameter of 350 μm. The pressure drop values across the bed were recorded for different superficial gas velocities (0.05 – 0.40 m/s). The minimum fluidization velocity obtained by plotting the pressure drops against the superficial gas velocity, is about 0.15 m/s. This result shows that the theoretical minimum fluidization velocity specified in Table 1 for the bed, is about 14% lower than the experimental value.

Table 1. Parameters for model computations.

Parameters	Values	Units
Particle diameter, d_p	350	μm
Particle sphericity, $\emptyset_s$	1.0	-
Particle density, ρ_s	2500	kg/m^3
Gas density, ρ_g	1.186	kg/m^3
Gas viscosity, μ_g	1.78x10^{-5}	Pa.s
Gas constant, R	0.287	kJ/(kg-K)
Gas temperature, T	25	^{0}C
Gas reference pressure, P_{ref}	1.0	bar
Initial bed height, h_0	0.32	m
Initial solid volume fraction, ε_0	0.52	-
Minimum fluidization velocity (19), U_{mf}	0.129	m/s
Bed height at minimum fluidization, h_{mf}	0.32	m
Superficial gas velocity, U_0	0.05 – 0.40	m/s
Maximum solid volume fraction, ε_{smax}	0.63	-
Restitution coefficient, e	0.90	-
Simulation time step	0.001	s
No of cells	125	-

Table 2. Initial and boundary conditions.

Initial Conditions	$0 \le z \le h_0$ $p_g(0, z) = \varepsilon_0 \rho_s g(h_0 - z)$ $\varepsilon_s(0, z) = \varepsilon_0$ $h_0 < z \le H$ $P_g(0, z) = 0$ $\varepsilon_s(0, z) = 0$ $0 \le z \le H$ $u(0, z) = U_0/\varepsilon_g$ $v(0, z) = 0$
Inlet Boundary	$u(t, 0) = U_0$ $v(t, 0) = 0$ $p_g(t, 0) = \varepsilon_0 \rho_s g h_0$
Outlet Boundary	$p_g(t, H) = 0$ $\varepsilon_s(t, H) = 0$

Figure 2 compares the simulated fluid pressure with the experimental data. The simulated results are obtained from the 1D model presented here and a 3D model reported by Thapa and Halvorsen (2013). As can be seen, the simulated data agree well with the experimental results at a height of 0.13 m above the distributor. At this height, the predictions from the 1D-model are better compared with the predictions from the 3D models. At the height of 0.23 m, the 1D model results also agree very well with the results from the 3D model.

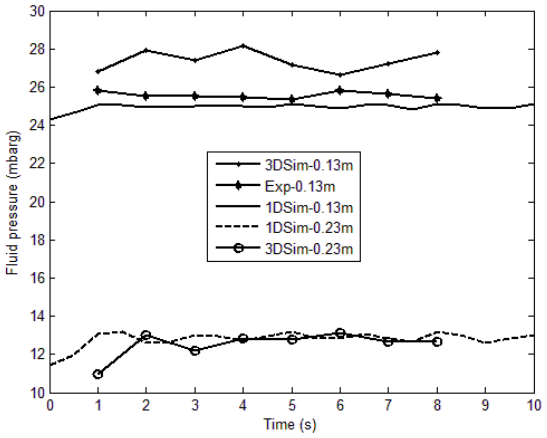


Figure 2. Evolution of fluid pressure at superficial gas velocity 0.18 m/s.

Figure 3 shows the time-averaged velocities of the fluid and particles for two different superficial gas velocities, 0.18 m/s and 0.32 m/s. From these results, it can be seen that the fluid velocity at the exit of the column is slightly higher than the velocities at the inlet. This variation in the fluid velocity along the bed axis is probably due to changes in the fluid density along the bed height. The figure also shows that fluid velocities within the bed are higher than the inlet velocities, which could be due to lower flow area available for the gas as particles occupy space within this region. The variation of particle velocity within the bed at different gas velocities conforms to the solid movement pattern described by Kunii and Levenspiel (1991). Figure 4 gives the instantaneous solid volume fractions for the respective velocities after 5, 10 and 20 s. These results show that the movement of particles in the fluidized bed are more vigorous with higher superficial gas velocity.

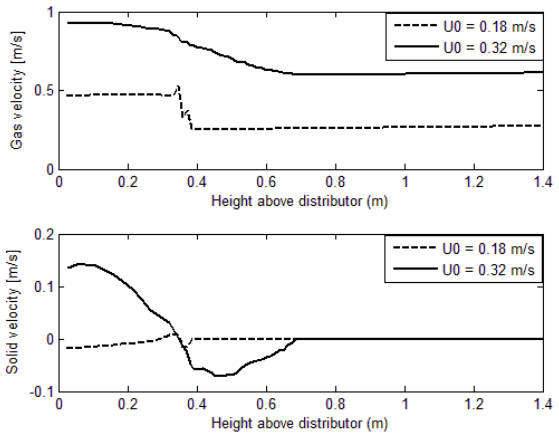


Figure 3. Time-averaged velocity profiles for fluid (upper plot) and particles (lower plot).

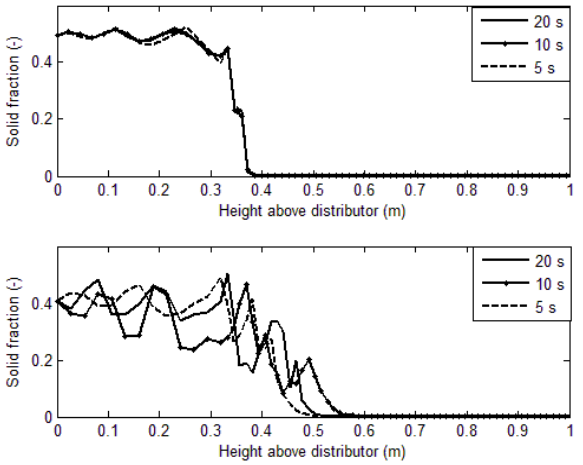


Figure 4. Instantaneous profile of solid fraction with superficial velocities 0.18 m/s (upper plot) and 0.32 m/s (lower plot).

The variation of average void fraction with superficial gas velocity within the dense region is shown in Figure 5. The average void fraction is obtained up to the height of 0.32 m above the distributor. The figure shows that the void fraction increases with increasing superficial gas velocity. It can also be seen that the bed transits into different regimes within different ranges of the superficial gas velocity.

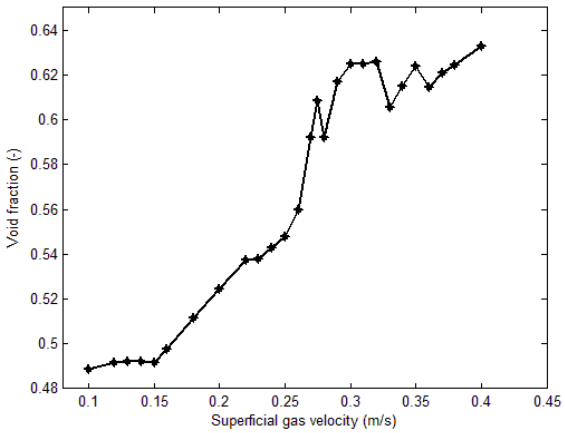


Figure 5. Variation of average bed void fraction with superficial velocity.

Four different flow regimes can be distinguished from Figure 5. Below 0.14 m/s, the bed's void fraction is about 0.49. Within this region, the bed behaves like a fixed bed with all the particles retained within the dense bed. The abrupt increase in the void fraction after 0.14 m/s indicates that the bed is fluidized. As expected for a Geldart B solid, the bed will begin to bubble when the velocity is above 0.14 m/s. Between 0.14 m/s and 0.22 m/s, the void fraction increases linearly. Beyond 0.22 m/s, it increases exponentially with an increase in the gas velocity up to 0.27 m/s. Within this velocity, the bed is more agitated with fast-rising bubbles. From (23), the minimum gas velocity for the onset of slugging is about 0.26 m/s. Since $\frac{h_0}{D_h} = 3.81\ (> 2)$, there is possibility of slug flow in the bed when the superficial gas velocity is above 0.26 m/s. From Figure 5, it can be seen that the void fraction flattens out with a superficial gas velocity beyond 0.27 m/s. More so, the variation of void fraction above 0.27 m/s fluctuates as the gas velocity increases, which shows that the bed is slugging. Thus, the velocity 0.27 m/s is the gas velocity at onset of slugging based on this simulation. The fluctuation of the bed void fraction as the velocity increases could be because in a slug flow the bed does not have a clear defined height over which the averaging is taken. In comparison, similar phase changes have been experimentally observed in Sundaresan (2003) with beds of fine particles that can readily agglomerate. With the simulated minimum fluidization velocity being 0.14 m/s, compared with the experimental value of 0.15 m/s, and with the simulated gas velocity being 0.27 m/s compared with the theoretical value of 0.26 m/s at onset of slugging, it can be concluded that the 1D model predicts the bed flow behaviour reasonably well.

Figure 6 shows the profiles of solid volume fraction at velocities 0.08, 0.20, 0.25 and 0.32 m/s, hence comparing the different flow regimes shown in Figure 5. The result shows that within the bubbling regime, the bed height expands by about 0.04 m (representing 12.5%) above the height at the minimum fluidization. The decrease in the solids fraction as the gas velocity increases is accompanied with a small fraction of particles in the freeboard. This keeps the mass of particles in the column balanced. In the solid region (fixed bed), the bed height is reduced below the settling bed height (about 0.32 m, accompanied with an increase in solid fraction), owing to the fact that the bed is closely packed towards the maximum packing solid fraction of about 0.63 used in the simulation. The figure shows that in the slugging regime, the bed expands unevenly with some particles flowing into the freeboard up to a height of 0.6 m. This shows that the bed height is not clearly defined within the slugging region, as can also be seen from Figure 7.

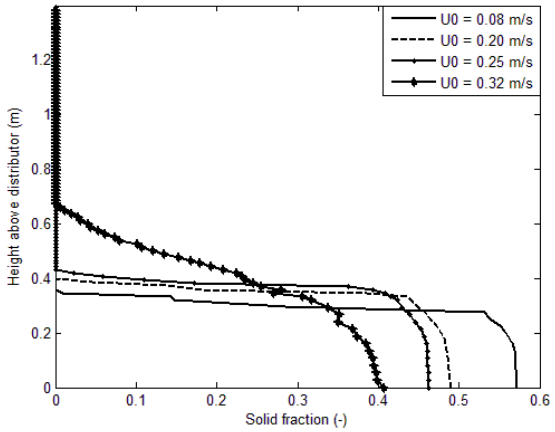


Figure 6. Simulated profile of solid fraction at different bed flow regimes.

Figure 7 shows the profile of solid fraction for some velocities within the slugging regime. Within the height interval 0.2 – 0.4 m, the solid fractions at velocity 0.32 m/s are lower than the corresponding solid fractions at velocity 0.36 m/s. This explains why the void fraction fluctuates with increasing superficial gas velocity within the slugging regime as given in Figure 5. Figure 7 also shows that the average solids volume fraction for the same range of velocities within the slugging regime is almost the same.

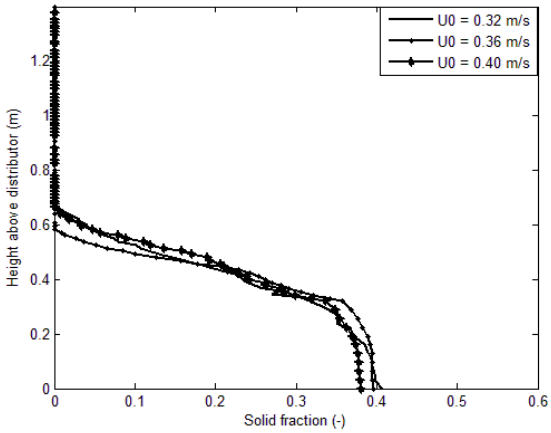


Figure 7. Simulated profile of solid fraction at different velocities within slugging regime.

6 Conclusions

This paper presents a detailed one-dimensional model based on the Euler-Euler approach for predicting hydrodynamics of a bubbling fluidized bed. The solution algorithm includes a void propagation equation that accounts for the effect of fluid density variations. The method developed here is computational efficient, taking only 10 minutes computer time for simulation of a 20 s flow in the bed, against several hours required in a 3D model computation.

Qualitatively, the results show that the 1D model predicts the different regimes of a fluidized bed. The simulated minimum fluidization velocity agrees well with the experimental data, and the value of gas velocity at the onset of slugging compares well with the value obtained from the empirical expression proposed by Geldart (1986).

Further work will include full validation of the 1D model against a 3D model results and analysis of sensitivity of the model to different parameters.

References

J. Baeyens and D. Geldart. An Investigation into Slugging Fluidized Beds. *Chemical Engineering Science*, 29: 255 – 265, 1974.

H. M. Benzarti and H. Bournot. Drag Models for Simulation Gas-Solid Flow in the Bubbling Fluidized Bed of FCC Particles. *International Journal of Chemical, Molecular, Nuclear, Materials and Metallurgical Engineering*, 6 (1), 2012.

C. T. Crowe, J. D. Schwarzkopf, M. Sommerfeld, and Y. Tsuji. *Multiphase Flows with Droplets and Particles*, 2nd ed., Taylor & Francis Group, Boca Raton London, New York, USA, 2012.

J. F. Davidson and D. Harrison. *Fluidized Particles*, Cambridge University Press, New York, 1965.

S. Ergun. Fluid Flow through Packed Column. *Chemical Engineering Progress*, 48: 89 – 94, 1952.

D. Geldart (Ed.). *Gas Fluidization Technology*, 1st ed., John Wiley & Sons, Ltd., pp. 53 – 97, Chap. 4, 1986.

D. Gidaspow. *Multiphase Flow and Fluidization*: Continuum and Kinetics Theory Descriptions, Academic Press Inc., San Diego, California, USA, 1994.

D. Kunii and O. Levenspiel. *Fluidization Engineering*, 2nd ed Butterworth – Heinemann, Washington Street, USA, 1991.

D. Lathowers and J. Bellan. Modeling of Dense Gas-Solid Reactive Mixtures Applied to Biomass Pyrolysis in a Fluidized Bed. In *Proceedings of the 2000 US, DOE Hydrogen Program Review*, NREL/CP-570-28890, USA, 2000.

P. Li, X. Lan, C. Xu, G. Wang, C. Lu and J.Gao. Drag Models for Simulating Gas-Solid Flow in the Turbulent Fluidization of FCC Particles. *Particuology*, 7: 269 – 277, 2009.

J.D. Silva. Numerical Modelling of the Fluid Dynamics in a Bubbling Fluidized Bed Biomass Gasifier. *Journal of Petroleum and Gas Engineering,* 3 (3): 35 – 40, 2012.

J. Solsvik, Z. Chao, and H. A. Jakobsen. Modeling and Simulation of Bubbling Fluidized Bed Reactors using a Dynamic One-dimensional Two-Fluid Model: The Sorption-Enhanced Steam-Methane Reforming Process. *Advances in Engineering Software,* 80: 156 – 173, 2015.

S. Sundaresan. Instabilities in Fluidized Beds. *Annual Review of Fluid Mechanics*, 35: 63 – 88, 2003.

M. Syamlal. *MFIX Documentation Numerical Technique*. Report, Department of Energy, Federal Energy Technology Center, DOE/MC 31346-5824, USA, 1998.

F. Taghipour, N. Ellis, and C. Wong. Experimental and Computational Study of Gas-Solid Fluidized Bed Hydrodynamics. *Chemical Engineering Science*, 60: 6857 – 6867, 2005.

R.K. Thapa and B.M. Halvorsen. Study of Flow Behaviour in Bubbling Fluidized Bed Biomass Gasification Reactor using CFD Simulation. In *Proceedings of the 14th International Conference on Fluidization – from Fundamentals to Products*, Eds. ECI Symposium Series, Volume, 2013.

C.Y. Wen and Y.H. Yu. A Generalized Method for Predicting the Minimum Fluidization Velocity. *AIChE J.*, 12: 610 – 612, 1966.

N. Xie, F. Battaglia, and S. Pannala. Effects of Using Two– Versus Three-Dimensional Computational Modeling of Fluidized Beds: Part I, Hydrodynamics. *Powder Technology*, 182: 1 – 13, 2008.

W.C. Yang (Ed.). *Handbook of Fluidization and Fluid-Particle Systems,* Marcel Dekker, Inc., 2003.

A New Concept of Functional Energetic Modelling and Simulation

Mert Mokukcu[1,2] Philippe Fiani[1] Sylvain Chavanne[1] Lahsen Ait Taleb[1]
Cristina Vlad[2] Emmanuel Godoy[2] Clément Fauvel[3]

[1]Sherpa Engineering, France, {m.mokukcu,p.fiani,s.chavanne,l.aittaleb}@sherpa-eng.com
[2]Automatic Control Department, Laboratoire des Signaux et Systèmes (L2S, UMR CNRS 8506) CentraleSupélec-CNRS-Université-Paris-Sud, France, {cristina.vlad,emmanuel.godoy}@centralesupelec.fr
[3] IRCCyN, UMR-CNRS 6597, Ecole des Mines de Nantes, IMT, France, clement.fauvel@mines-nantes.fr

Abstract

In this study a new concept of functional modelling and simulation is introduced. First, the necessity and the expected outcomes of the new concept are explained. Secondly, the methodology of functional modelling based on a modular concept and the basic elements are presented, with details of OFS (Organico Functional Set). Then, the implementation of the new modelling concept using Sherpa Engineering's PhiSim environment is described in order to perform simulations. Finally, the proposed modelling method is applied to two different applications: a generic parallel hybrid electric vehicle (HEV) and a waste water treatment unit of a building. Simulation results of parallel HEV application are also presented.

Keywords: electric vehicles, functional modelling, functional model simulation, hybrid vehicles, waste water treatment unit

1 Introduction

The economic and ecological context drives industries and academic research to investigate how systems can be optimally designed with respect to their local and global energy efficiency (Arnal et al., 2011; Mouhib et al., 2009; Sherpa Engineering, 2016; Wirasingha et al., 2011). Due to the quick progress and the variety of energy technologies and energy management strategies, being able to numerically simulate a solution has become a crucial aspect of the system design process (Haveman et al., 2015). Usually, this step requires a model of the system and a simulation environment. So far, the Bond Graph theory introduced by Karnopp and Rosenberg has been widely promoted in industrial and academic communities to model multi-physic systems (Karnopp et al.,1983).

Basically, this approach is based on effort and flow interactions and uses phenomenological analogies to represent any nature of systems (i.e. mechanics, electrics, etc.). The obtained model can subsequently be used for system analysis and design of optimal control laws (Bideaux et al., 2006; Otter et al., 1996). This philosophy forms the base of several commercial multi-physics modelling tools as AmeSim, Dymola or PhiSim (Marquis-Favre et al., 2006; Pénalva et al., 1994; Sciarretta et al., 2007). However, these tools allow only constructional design, i.e. the representation of the organic level of a system.

In his studies, (Von Bertalanffy, 1968) remarked that some systems, referred as complex systems, contain many interactions with themselves and their environment that should be designed with a unique level of abstraction (Eriksson, 1997; Le Moigne, 1994). In addition to constructional design, the use of functional approach in system modelling has been largely supported by most of complex systems specialists (Le Moigne's modelling theory (Eriksson, 1997; Le Moigne, 1994), Sagace methodology (Sciarretta et al., 2007) and axiomatic design (Suh, 1998)). This approach defines a higher abstraction level than for constructional design (see Figure 1) and does not need the organic elements definition. The system is modelled through functionalities which are interacting in order to achieve a certain purpose referred as system mission. Therefore, it can be used in early stage development to simulate the system architecture. However, the framework to be used (i.e. equations, variables flows) is currently an open problem, especially for energetic systems.

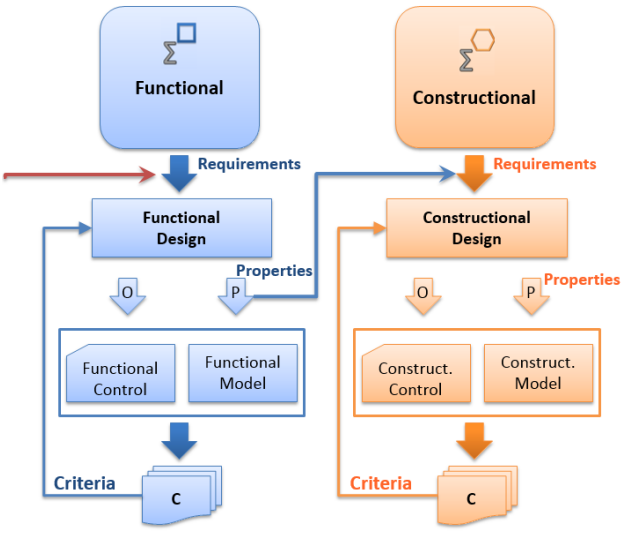


Figure 1. Functional and constructional (organic) modelling.

The difficulties encountered in these multiple level models are mainly the following: how to define the model and its parameters when the equipment does not yet exist, and how to avoid the adjustment of the entire energy supervision system (responsible for an optimal energy flow) at each modification of the architecture. These difficulties increase the time required to model and analyse the simulation results leading either to a reduced number of potential solutions or to a laborious design process. Confronted by these challenges, it has become essential to develop a new tool-based methodology, also based on modelling, that uses the necessary level of abstraction by integrating modular functional models and optimization algorithms.

A first step towards the formulation of a functional design language for energetic systems has been made with *PhiGraph* introduced in (Brunet et al.,2005). However, interactions consist of effort and flow exchanges, as in Bond Graph theory, and belong to the organic level of a system. Therefore, PhiGraph cannot be related to a full functional language. This is why a new concept of functional modelling language and method is proposed.

The expected outcomes of the proposed functional modelling method are:

- Fast simulation and evaluation of the system concept before choosing the technology;
- Simulation of the system as a whole: physics and control;
- Obtaining a supervisor for the organic simulation model;
- Make connections between modelling and simulation in multi-physic systems.

In our previous work (Fauvel, 2015), two functional elements were introduced: a *consumer* and a *source*, which are exchanging *needs* and *availability* information. This has led to a fully functional framework which was initially applied to formulate an energy management problem (Fauvel et al., 2014). This paper extends the concept of *consumers* and *sources* ports to a complete simulation language for energetic systems by considering five functional elements described in Section 2. Its potential is highlighted by modelling and simulating two classic applications in Section 3. Conclusion and perspectives of this work are given in Section 4.

2 Functional Modelling Methodology

The base concept of the proposed functional modelling method is to provide a functional link between two systems, which can be described as an exchange in terms of energy (*mechanical, electrical, hydraulic, thermal, etc.*), matter (*fluid, solid, etc.*) and information (*set point, measurement, etc.*). In the early steps of system design, this exchange and its nature have to be defined for two sub-systems or for a system and its environment.

In constructional or organic level of modelling like Bond Graph or PhiGraph, inputs and outputs of the components depend on flow/effort of physical domain. Unlike the organic level, the functional level uses three types of flow: energy, matter and information, each of them being decomposed in a triadic basis. In Table 1, the transformation natures of functional level are presented.

Table 1. Transformation Natures of Functional Modelling.

T	*Time Transformation*	Storage, Accumulation
S	*Space Transformation*	Transport, Transmission, Distribution, Injection, Extraction
F	*Form Transformation*	Transformation, Conversion, Production, Destruction, Consumption

2.1 Modular Concept

In a controlled complex system, the exchange between different functions is governed by a certain need. For example, in order to drive at a given speed, the vehicle motion functionality (*Mobility*) needs power flux. If the need is not expressed, there is no reason to supply power to *Mobility*. In the proposed methodology, the need has dominant causality with respect to the energy/matter supply. On this basis, the functional modelling method is developed from the following question: *Who transmits a need, and to whom?*

This methodology introduces two types of ports: source port, which supplies an energy/matter flux as an answer to an expressed need, and consumer port, which receives energy/matter flux also as an answer to a specified need. In Figure 2 the source and consumer ports between two systems are introduced. In order to highlight the dominance of a causality requirement, Figure 2 illustrates the functional link with a single arrow, which expresses the necessity of causality. The arrow direction indicates where the need is transmitted, regardless its value (positive/negative) or the direction of supply. Furthermore, at this level of abstraction, it is not necessary to specify the physical domain associated with the need, but it can be done for information.

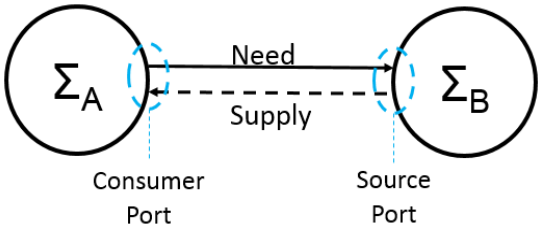


Figure 2. Consumer and source ports.

Figure 3 presents the functional links for a system with three sources and their associated source and consumer ports. The energy distribution in the system is made by distribution elements, represented by grey bars, and that will be introduced in Section 2.2.4.

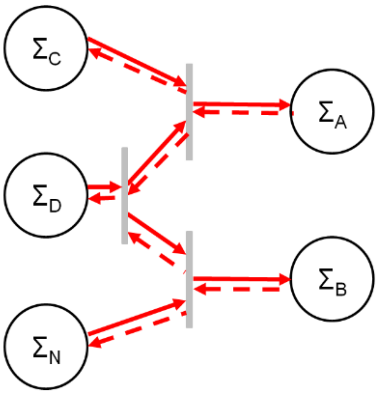


Figure 3. A system example with multiple source and consumer ports.

In the proposed concept, all elements use simple equations from these transformation natures. Some examples of equations can be introduced as: energy (1) and dynamics (2) of time transformation nature, efficiency (3) of form transformation nature and power balance (4) of space transformation nature:

$$E = \int P \tag{1}$$

$$Y(s) = H(s)U(s) \tag{2}$$

$$\rho = \rho(\mu, \dots) \tag{3}$$

$$\sum P_t \tag{4}$$

Equations are used to define internal properties of different elements: the integral of power is used for storage elements, transfer functions for effector elements, efficiency for transformation elements and power balance equation for distribution elements.

2.2 Adequate Language in Modelling Level

In this section, basic elements of functional modelling are described. These elements are: source, storage, transformation, distribution and effector, classified as in Table 2. Generally, all functional elements are based on two blocks: control system block (upper block) and operating system block (lower block), usually represented respectively in green and blue colors as shown in Figure 4(a).

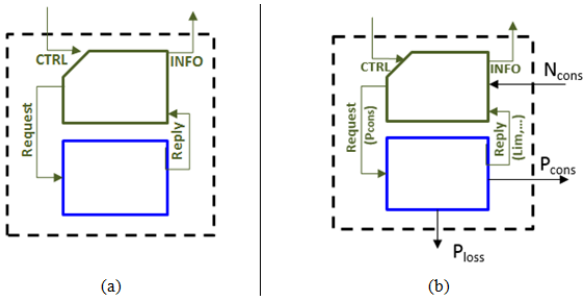


Figure 4. Representations of (a) generic and (b) source elements.

A generic element is a black box that has its own control and operating systems. The control system manages the operation according to the demands, actions and constraints. On the other hand, the operating system represents the physical behaviour of the function.

Table 2. Element Types of Functional Energetic Modelling.

Source	*Storage*	*Transformation*	*Distribution*	*Effector*
Energy & Matter Source	Energy & Matter Storage	Energy & Matter Transformation in Different Domains	Energy & Matter Distribution	Represents Energetics Services

2.2.1 Source

The source (e.g. fuel station, electrical grid) represents the supply of energy/matter to meet the consumer needs. Within the physical limits of the source, the control system is intended to compute the provided power P_{cons} in response to the received need N_{cons}, illustrated in Figure 4(b). If losses are considered, the source also contains P_{loss} port.

As a consequence of many possible sources, the operating system behaviour is not generic. In the simplest case, the source is infinite like an electrical grid, where the only feature that defines the maximum power is the received or the supplied power. In a more specific case, such as a brake system, which can be represented by a negative power source, the maximum power that can be dissipated depends on its organic characteristics (maximum torque etc.).

2.2.2 Storage

A storage (e.g. battery, fuel tank) represents a given energy/matter storage which engages availability for an answer (P_{cons}) to a need of consumption (N_{cons}) and for its own need (N_{charge}) to charge (P_{charge}), to keep its state of charge (SOC) at an adequate level. If necessary, the settings of this block can be customized for a particular type of storage. For example, the maximum energy that can be stored in a fuel tank is linked to its volume, which does not exist for a battery.

In Figure 5(a), the representation of the storage elements is given, which has a simple operating system behavior. The stored energy/matter level is achieved by integrating the balance of incoming-outgoing powers, while considering the capacity limitations of the storage

(the maximum power that can be received or supplied). Exceeding these limitations may result in losses (P_{loss}). The equation adopted to represent the behavior of the operating system is quasi-generic.

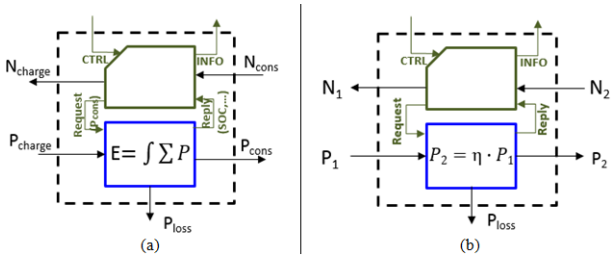


Figure 5. Representations of (a) storage and (b) transformation elements.

2.2.3 Transformation

The transformation element (e.g. electric machine, internal combustion engine (ICE), converter) offers need and power transfer between two functional elements. It takes given efficiency into consideration and it also allows domain change (e.g. fuel to mechanical). Figure 5(b) illustrates the representation of the transformation element.

The control system of transformation converts the received need (N_2) into need delivered (N_1) with a specified efficiency. Transmitted power (P_2) is derived from the received power (P_1) regarding any limitations. The difference between P_1 and P_2 ports defines the power losses. The transformation element is also characterized by a quasi-generic equation.

2.2.4 Distribution

The distribution element of functional modelling can be seen as a connector of multiple sources and consumers. This element has two main tasks: distribution of consumer needs (N_k, N_l, N_m) to sources (N_i, N_j) and distribution of supplied power/matter (P_i, P_j) to consumers (P_k, P_l, P_m). Distributions are allowed by taking into account the constraints specific to each source and consumer. The representation of distribution is given in Figure 6(a). Distributions respect the balance equation of energy/matter, based on the principle that there are no losses or storage in normal operation. Moreover, they connect the ports of same nature.

The distribution control system is more complex because of management of multi-sources/multi-consumers. The designer's task is to choose the most appropriate algorithm that will assure optimal power distribution with respect to system requirements and physical limits. The distribution constraints are related to the source availability, the consumer priorities and the evolving distribution method.

2.2.5 Effector

The effector element is associated with the achievement of an objective. In order to achieve its objective, the effector transmits a power need (N_{cons}). Accordingly, it receives power (P_{cons}) to execute its function. Figure 6(b) indicates the need generation and the power reception.

The effector is the heart of the functional modelling architecture since it generates the need. Without effector there is no need, therefore no functional architecture design can be made. The control system is intended to compute the power need to achieve the objective. Furthermore, the operating system is not generic because of the great variety of its objective and execution (i.e. for thermal comfort, temperature value can be obtained from heat equation).

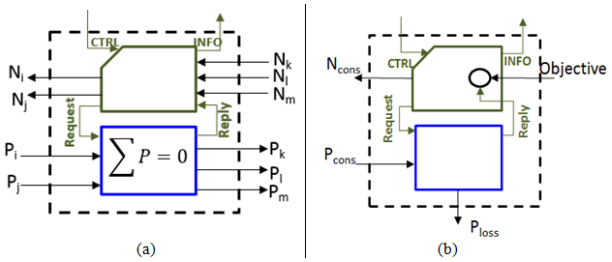


Figure 6. Representations of (a) distribution and (b) effector elements.

2.3 Details of Functional Energetic Modelling Method

OFS is a generic element (Figure 7(a)) that can be used to represent all five elements introduced in the previous section. For example, a vehicle can be represented as two interconnected OFS, one to represent fuel source and a second model for mobility effector (Figure 7(b)).

Generally, the input arrow is connected to a source port of OFS and the output arrow is connected to a consumer port of OFS. If wanted, arrow colours could be used as domain instructors as red for electric, green for mechanic and blue for hydraulic. At the top of the element, the green box offers an information link between the OFS and the supervisor, which is a global controller of the system. Finally, the ground symbol placed at the bottom of the representation corresponds to losses that will not be recycled. To recover losses, an additional source port could be used.

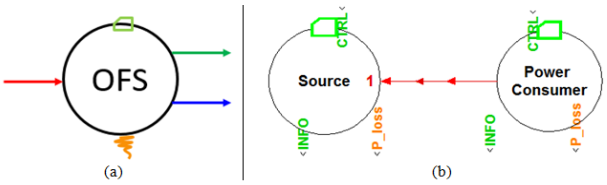


Figure 7. (a) Representation of OFS and (b) example PhiSim link.

3 Implementation to Simulation Environment

This section describes the implementation of energetic and functional modelling elements in PhiSim, an environment developed by Sherpa Engineering to define

and simulate the proposed functional model. Sherpa Engineering proposes PhiSim as a modelling and simulation environment for physical systems using Matlab/Simulink software that allows generating models and their control in a multiport environment (Sherpa Engineering, 2016).

3.1 Simulation Tool Integration

Two different types of standardized ports are considered: source port and consumer port. Source port receives consumer power need, availability, acceptances of power and energy/matter from connected consumer. The source port also transmits provided power, availability, acceptances of power and energy/matter to the connected consumer.

The elements are connected based on the following principles:

- Communication between elements and supervisor is possible using a bidirectional link of control/information (CTRL/INFO);
- Two elements are interconnected by a bidirectional link of need/supply (N_*, P_*);
- Losses port of an element is defined as an output port that does not require an input (it refers to losses without an associated need);
- The direction of arrows indicates where the need is transmitted regardless its sign.

Availability and acceptancy information are useful for the (local or global) control. For example, to distribute a need of a consumer to multiple sources, distribution must check the sources availability. For instance, a consumer is available to receive negative power which is provided by a source that has acceptance.

It is important to highlight the role of the distribution element for the intelligence system and the energy management strategy. For the moment, power need for each consumer is treated using priorities. Total power need is spread across different sources and prioritized according to their availability and acceptance. If total power need cannot be fully allocated, the remaining power need is allocated to the source that has first priority.

3.2 Examples from Different Domains

The functional modelling methodology presented in the previous sections is applied to two complex systems from different fields of applications, showing the general character of the proposed modelling formalism and the capacity to easily adapt to various systems. The considered applications are: a parallel hybrid electric vehicle (HEV) and a waste water treatment unit of a building.

For the first example, the objective is to evaluate the consumption of a parallel HEV with respect to a specific power load profile. In order to analyze the power flow in the system, the vehicle dynamics has to be considered, as well as the electrical auxiliaries and the braking system. All these elements lead to a complex organic model which will increase the simulation time and the design procedure: architecture choice, supervisor construction, component choice and sizing.

Using the functional modelling approach, the elements of the parallel HEV architecture are represented by elements described in Section 2.2. The implementation of the functional model in PhiSim is given in Figure 8.

A standard driving cycle, NEDC (New European Driving Cycle), is chosen to analyse the architecture of the parallel HEV using a functional model that can be simulated.

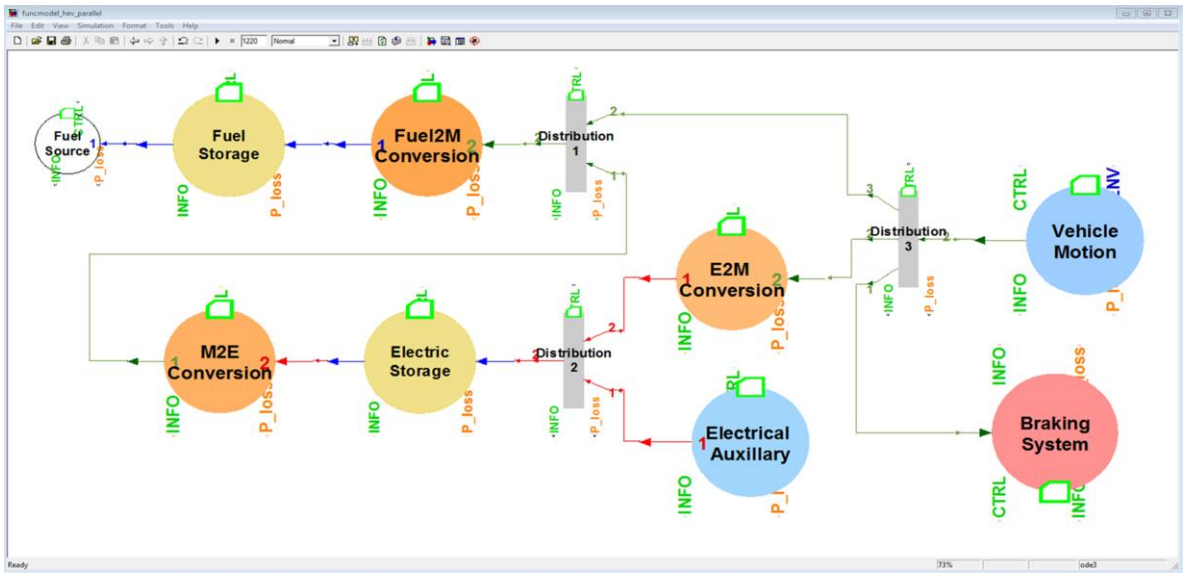


Figure 8. PhiSim functional model of a parallel HEV.

The simulation model of the hybrid vehicle consists of:

- Three energy distribution elements (Distribution),
- Three transformation elements (F2M, M2E and E2M),
- Two storage elements (Fuel and Electric),
- Two effector elements (Electrical auxiliary and Vehicle motion).

A special functional element is employed for the braking system, which can act as a source or an effector for different driving modes. The simulation results of this application are given in Section 3.3.

The second complex system is a waste water treatment unit of a building. The objective of this application is to calculate the cold and hot water consumption for a real scenario and also to provide the equivalent power consumption and cost estimation. The difference with the parallel HEV is in the flow type used: matter and energy instead of energy.

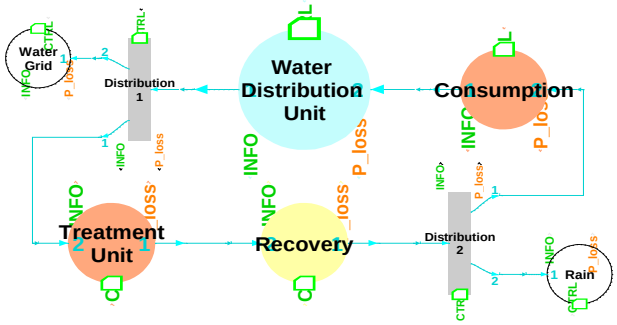


Figure 9. Waste water treatment unit functional model.

In this case, the functional model is obtained based on a reduced organic model with a real scenario of grey water treatment unit (GWTU) from a hotel. The functional model of the waste treatment unit is given in Figure 9 and Figure 10.

The model consists of the following elements:

- Two transformation elements (Treatment unit and Consumption),
- A storage element (Recovery as water tank),
- Two distribution elements,
- Two source elements (Rain and Water grid),
- A system of water distribution unit (see Figure 10) of a building.

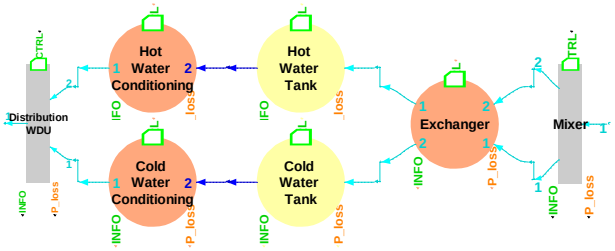


Figure 10. Waste water distribution unit functional model.

The water distribution unit is modelled by:

- Three transformation elements (Hot and Cold water conditioning and Exchanger),
- Two storage elements (Hot and Cold water tanks),
- Two distribution elements (Mixer and Water distribution unit).

As it can be seen from both examples, functional modelling representation and simulation can be applied to different domains and natures. For example, for HEV application the flow nature is either energy or information, whereas in GWTU application, in addition to energy, matter flow is used. However, information is always used as a flow nature in order to have a supervisor that manages the energy flow of the designed system. Therefore, the functional model of the system is thought to be used as a supervisor of the organic model.

3.3 Parallel HEV Example Simulation Results

The simulation results of the parallel HEV are represented in Figure 11. As equations used in the proposed functional modelling method are quite simple, the design time is significantly reduced for a preliminary analysis and a full driving cycle can be simulated in a few seconds.

- The first figure (Figure 11(a)) shows that the speed of the parallel HEV is consistent with the NEDC profile.
- The result can be validated as vehicle power needs (Figure 11(b)) and source power supplies (Figure 11(c)) are appropriate to real time calculations.
- In Figure 11(d), the scaled values of the SOC (state of charge) are presented. Electrical storage and fuel storage SOC can be analysed using the functional model, as well as the sources powers.

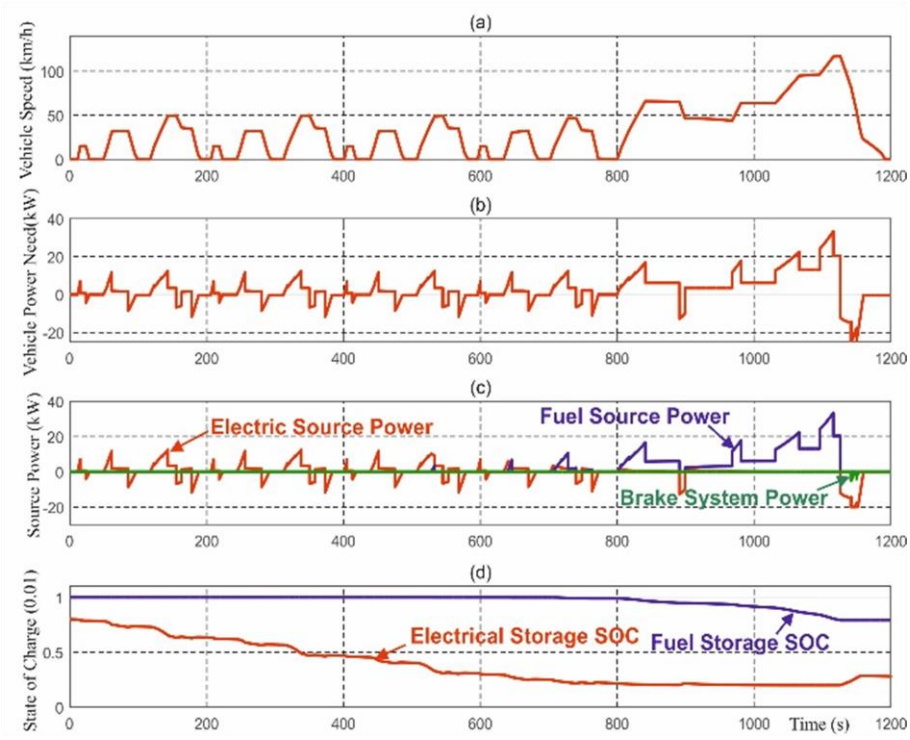


Figure 11. Functional model simulation of a parallel HEV.

These results can give a head start for choosing the system architecture and lead to an organic modelling and its simulation. As mentioned in Section 3.2, the proposed modelling method leads to a supervisor/controller of the organic model with adjustments to its flow natures. For example, for the electrical-to-mechanical transformation element, the functional modelling flow is power. In the organic modelling level, this flow becomes electrical flow or mechanical rotation flow. Thus, a transformation element that adjusts the command flows will be added to the system for the supervisor.

4 Conclusions and Perspectives

A concept of functional modelling is proposed and applied to different complex systems where the flow nature is either energy or matter or both. The functional modelling approach develops a macro model of a complex system that can be easily and quickly simulated using the simulation framework PhiSim.

A short-term perspective of this work is to apply this modelling concept to systems with both energy and matter flows for simulations. The long-term perspective is to improve the intelligence of distributions. So far, distributions use priorities for need and power distribution. The improvement consists in developing a performant algorithm for need/supply distributions capable to optimize all natures of flows, and fast enough in order to minimize the simulation time. For these reasons, the proposed modelling formalism represents an interesting solution for industrial applications as it allows obtaining relevant results in a first stage of preliminary analysis of the system.

References

E. Arnal, C. Anthierens, and E. Bideaux. Consideration of glare from daylight in the control of the luminous atmosphere in buildings. *IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), Budapest, Hungary*, pages 1070–1075, 2011. doi: 10.1109/AIM.2011.6027070.

E. Bideaux, J. Laffite, W. Marquis-Favre, S. Scavarda, and F. Guillemard. System design using an inverse approach: Application to the hybrid vehicle powertrain. *Journal Européen des Systèmes Automatisés (JESA)*, Lavoisier, 40(3):269–290, 2006. doi: 10.3166/jesa.40.269-290.

J. Brunet, L. Flambard, and A. Yazman. A hardware in the loop (HIL) model development and implementation methodology and support tools for testing and validating car engine electronic control unit (ECU). *International Conference on Simulation Based Engineering and Studies, TCN CAE, Lecce, Italy*, 2005.

D. Eriksson. A principal exposition of Jean-Louis Le Moigne's systematic theory. *Cybernetics and Human Knowing*, 4(2-3):33–77, 1997.

C. Fauvel. Approche modulaire de l'optimisation des flux de puissance multi-sources et multi-clients, à visé temps reel. *Automatique/ Robotique, Ecole des Mînes de Nantes*, 2015. doi: tel-01245429.

C. Fauvel, F. Claveau, and P. Chevrel. Energy management in multi-consumers multi-sources system: A practical framework. In Proceedings of the 19th *IFAC World Congress, Cape Town, South Africa*, volume 47, pages 2260–2266, 2014. doi: http://dx.doi.org/10.3182/20140824-6-ZA-1003.02446.

S. P. Haveman and G. M. Bonnema. Communication of simulation and modelling activities in early systems engineering. *Procedia Computer Science.* Volume 44, pages 305–314, 2015. doi: 10.1016/j.procs.2015.03.021.

D. Karnopp and R. Rosenberg. *Introduction to Physical System Dynamics.* McGraw-Hill, 1983.

J-L. Le Moigne. *La théorie du système général, théorie de la modélisation*, 1994. Edited by ae Mcx, 2006. Available via http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:No+Title#0.

W. Marquis-Favre, E. Bideaux, and S. Scavarda. A planar mechanical library in the AMESim simulation software. Part II: Library composition and illustrative example. *Simulation Modelling Practice and Theory*, 14(2):95–111, 2006. doi: 10.1016/j.simpat.2005.02.007.

O. Mouhib, A. Jardin, W. Marquis-Favre, E. Bideaux, and D. Thomasset. Optimal control problem in bond graph formalism. *Simulation Modelling Practice and Theory*. Elsevier B.V., 17(1):240–256, 2009. doi: 10.1016/j.simpat.2008.04.011.

M. Otter, H. Elmqvist, and F. E. Cellier. Modeling of multibody systems with the object-oriented modeling language Dymola. *Nonlinear Dynamics*, 9(1–2):91–112, 1996. doi: 10.1007/BF01833295.

J. M. Pénalva and E. Page. SAGACE: la modélisation des systèmes dont la maîtrise est complexe. in *ILCE'94, Montpellier, France*, 1994.

A. Sciarretta and L. Guzzella. Control of hybrid electric vehicles, *IEEE Control Systems*, 27(2):60–70, 2007. doi: 10.1109/MCS.2007.338280.

Sherpa Engineering. *Phisim Brochure*, Paris, France. Available at: www.sherpa-eng.com. [accessed January, 2016].

N. P. Suh. Axiomatic Design Theory for Systems. *Research in Engineering Design*, 10(4):189–209, 1998. doi: 10.1007/s001639870001.

L. Von Bertalanffy. General System Theory. *Georg. Braziller New York*, 1968. Available via: http://books.google.es/books?id=N6k2mILtPYIC.

S. G. Wirasingha and A. Emadi. Classification and review of control strategies for plug-in hybrid electric vehicles. *IEEE Transactions on Vehicular Technology*, 60(1):111–122, 2011. doi: 10.1109/TVT.2010.2090178.

Taking Into Account Workers' Fatigue in Production Tasks: A Combined Simulation Framework

Aicha Ferjani[1] Henri Pierreval[1] Denis Gien[1] Sabeur Elkosantini[2]

[1] LIMOS, UMR 6158, SIGMA-Clermont, Aubière, France,
{aicha.ferjani,henir.pierreval,denis.gien}@sigma-clermont.fr
[2] Industrial Engineering Department, College of Engineering Riyadh, Kingdom of Saudi Arabia,
Selkosantini@ksu.edu

Abstract

In manufacturing systems, workers are often subjected to arduous working conditions, such as heavy loads and discomfort postures, which induce fatigue. Because of the effect of fatigue on workers' well-being, as well as on their performances, managers would need to understand the evolution of operators' fatigue during their work, in order to make relevant decisions (e.g. work schedule, facility layout decisions, and rest periods). In this context, we present a simulation modeling framework to evaluate manufacturing systems, which takes the workers' fatigue into account. The suggested framework combines several worldviews: Discrete Event modeling, multi-agent and System Dynamics. Discrete Event concepts are used to describe the manufacturing system dynamic behavior and agents are used to model workers. One important characteristic of agents' behavior on which emphasis is put is fatigue, which is modeled using System Dynamics concepts. The proposed approach is implemented using the Anylogic simulation software.

Keywords: simulation discrete event, intelligent agents, system dynamics, manufacturing systems, fatigue

1 Introduction

Workers play a key role in modern manufacturing systems. They are often subjected to arduous working conditions, such as carrying heavy loads, noise and vibrations. This leads to fatigue in work, which causes a decline in workers' performance, errors and may lead to health troubles. As consequence, when they have to organize a manufacturing system, managers are also interested in management policies that can contribute to improve workers' performance, as well as their wellbeing at work. In this respect, managers need to understand the evolution of operators' fatigue during their work, depending on the organizational scenario considered for the manufacturing system. It has been widely demonstrated that fatigue in manufacturing depends on several factors, such as the penibility resulting from the production environment (e.g. noise), and the amount of work carried out by each operator. This amount of work depends on several operating conditions in the factory, such as the schedule of breaks and the penibility of the tasks that workers perform. The assignment of the workers to the different machines in the system has an effect on the work duration of each operator. The evolution of workers' fatigue depends on how they will be solicited, which can have different impacts, in terms of system performance (e.g. delay and quality). Therefore, it is important to determine the production scenarios that induce excessive fatigue, in order to improve workers' performances and well-being at work. As consequence, there is a need for dynamic models that allow the evolution of fatigue along time to be evaluated, as well as the manufacturing system performance under different production scenarios. For this purpose, simulation seems to be a relevant approach. However, how to build a suitable model of manufacturing system that takes fatigue into account turns out to be a difficult research question.

To address this question, we propose a simulation modeling framework to model manufacturing systems that can take workers' fatigue into account. The suggested framework combines the paradigms of Discrete Event Simulation (DES), System Dynamics (SD) and Intelligent Agents (IA) in the same simulation model. The DES concepts are used to describe the dynamic behavior of the manufacturing system. Workers are modeled as intelligent agents with a specific knowledge and behavior. Since fatigue represents an important characteristic of this behavior, we use SD concepts to describe the evolution of the operators' fatigue during the work. The suggested combined simulation framework is implemented using the Anylogic simulation software, which allows these different worldviews to be combined.

The remainder of this paper is organized as follows. Section 2 explains the novelties of this study with respect to the existing literature. Section 3 focuses on the proposed framework. In Section 4, a first implementation of the suggested approach using the Anylogic simulation software is presented. The final

section draws the conclusions and some suggestions for future research.

2 Related research

In the last decades, researchers from different fields, such as the work psychology and ergonomics, have paid much attention to the fatigue phenomenon in manufacturing systems. In this context, few research works have used simulation to address such problems.

Digiesi et al. (2006 and 2009) have used simulation in order to investigate the impact of fatigue on the mean flowtime and the work in process in a manufacturing system. The authors have used DES to model a flow line system composed of one worker performing a repetitive task. However, fatigue is a continuous and a complex phenomenon that depends on several factors. For instance, Grandjean (1979) has highlighted that difficult tasks caused fatigue in manufacturing. Kahya (2007) has demonstrated that physical efforts and the penibility of the production environment contribute to increase the workers' fatigue.

Walters et al. (2000) have used DES to analyze the impact of fatigue and rotation schedules on workers performances. DES has also been used by Perez et al. (2014) to evaluate workers performances and cumulative spinal load in an assembly line. In these studies, the DES model of the manufacturing system is used to collect the data necessary (e.g. task durations, process time, and the availability of machines) for the work-rest models that they used to evaluate fatigue.

In the ergonomic literature, few approaches based on DES have been developed to help designers to understand the ergonomic impacts of a proposed alternative in system design. For example, Perez et al. (2014) have combined DES with a work-rest model to predict the effect of the mechanical exposure of workers and the accumulation of muscular fatigue. The authors have evaluated the accumulation of fatigue before and after performing a job such that, the time pattern of the cycle is given by the DES model of an assembly line. Thus, their model is not able to describe the evolution of fatigue during processing a task. Dode et al. (2015) have suggested an approach integrating fatigue-recovery pattern and learning into a DES model of an electronic assembly line to evaluate productivity and quality. The authors are interested in the muscular fatigue caused by the repetitive work. Therefore, they have evaluated fatigue through Muscular Endurance Time (MET) models, which are used when fatigue is assumed to be caused by the repetitive work.

According to Elkosantini and Gien (2009) workers are often assimilated to a simple resource with a failure rate, mean-time-between-failure and a repair time in most simulation models. However, workers, as human beings, exhibit a more complex and unpredictable behavior than the inanimate resources (Elkosantini and Gien, 2007). Therefore, authors as Boudreau et al. (2003) have emphasized the importance of modeling the human behavior of workers.

In this respect, Elkosantini and Gien (2009) have suggested an Agent-based framework to model the human behavior including different behavioral aspects, such as stress and satisfaction, in order to analyze the behaviors of workers in manufacturing. The authors do not model the manufacturing system in their approach.

Fatigue represents an important characteristic of the workers' behavior. As mentioned above, it depends on several factors, which may be complex and interact with each other. Thus, it is difficult to describe how this phenomenon evolves along time. In order to deal with this problem, SD concepts seem to be relevant. The SD approach (Forrester, 1958), has been widely used to understand complex phenomenon in different fields but it seems not to be used for fatigue models.

In summarizing, DES has been used by few researchers to address very specific problems. Certain concepts, such as IA and SD can be used in order to better model workers in simulation so as to take their behavior and fatigue into account. Although different concepts can be combined in simulation (Elkosantini and Gien, 2009), to the best of our knowledge, no publication seems to combine the concept of DES, SD and IA together in the same simulation model so as to evaluate the impact of different production scenarios on both system performance and workers' fatigue.

3 Related research Description of the proposed framework

3.1 Principle

To evaluate the impacts of the production scenario on the system performance and the workers' fatigue, we need to a simulation model of the manufacturing system in which the workers are subjected to fatigue. To meet this objective, we propose to combine different worldviews in the same simulation model. The proposed simulation-modeling framework combines the following paradigms: IA, DES and SD. For further information about the combined simulation, we refer to the review of Dessouky and Roberts (1997).

The workers' fatigue depends on their activities. For instance, when the worker performs a tiring job, his/her fatigue increases. However, fatigue decreases when the worker is resting. This leads to a dynamic behavior of the worker. According to Elkosantini and Gien (2009), the workers exhibit a dynamic and more complex behavior than the inanimate resources. Therefore, modeling workers as inanimate resources, as they are often modeled in the existing simulation models, seems to be not relevant to take their behaviors into account.

As defined by Franklin and Graesser (1996), an agent "continuously performs three functions: perception of

dynamic conditions in the environment; action to affect conditions in the environment; and reasoning to interpret perceptions, solve problems, draw inferences, and determine actions". According to Padgham and Lambrix (2000), the agent also shows a behavior. Based on the above definitions, we suggest using the concept of "Intelligent Agent" to model workers in our framework.

As mentioned in the related research, the behavior of workers in manufacturing is related to several aspects, such as the evolution of his/her fatigue or satisfaction and how to choose the tasks to be performed. These behavioral aspects are often complex. In order to describe the evolution of such behavioral aspects, we propose to use SD concepts, which are extensively used in the literature to describe the evolution of complex and continuous systems. Hence, a SD component describing how the worker behaves is integrated into the agent component of the model.

In order to describe how the worker behaves, we need to know several information about his/her work and the production environment. For instance, how many machines in the manufacturing and what is the physical effort required to work on each one. In addition, one needs to know the amount of work carried out by the agent to describe the impact of the accumulation of work on its fatigue. This requires information about the task completed, the work schedule and rest periods. This information is available thanks to the DES component of the manufacturing system model.

Figure 1 illustrates the global architecture of the suggested combined model. As it will be depicted below, the three components interact with each other.

3.2 Discrete Event Simulation component

The DES component represents the set of the elements of the manufacturing system (e.g. machines, products, transporters and stocks), which interact with the workers and, at the same time, have an impact on their fatigue. Thus, our DES component can describe the products, the manufacturing resources and several characteristics of the manufacturing system. The more classical performance criteria are computed in the DES component (e. g. mean flowtime of jobs (Neumann et al., 2006).

Typically, the machines in the system are modeled as resources, buffers are modeled as queues, etc. The penibility associated with the use of the machines has a great impact on the workers' fatigue. Thus, a coefficient Penibility of Machine is associated to each machine in the simulation model to describe the physical effort required by the worker to use that machine. We also need the position of each production facility since the distance walked by the operator also affects his/her fatigue. For that, each element in the DES component is characterized by its geographical coordinates.

The state variables describe the state of each element in the DES component. The dynamic behavior of the DES component has a significant impact on the amount of work that should be carried out by each worker. As consequence, the behavior of the system also affects the workers' fatigue. Therefore, we find that the state variables describing the state of each element in the DES component as well as, the events related to the start or the end of a production operation or transport activity, are used to evaluate the operators' fatigue.

As mentioned in the related research, the penibility of the production environment, such as the noise, temperature and vibrations, has a significant impact on the workers' fatigue. Thus, a subjective evaluation of the penibility of the production environment may be described, in our DES component, through the coefficient Penibility of Environment.

3.3 Agent model

As presented in Figure 1, we consider that the agent in our framework is characterized by:

- **Goals:** they can be classified in two types: the first type is common for all the agent, which is mainly related to the improvement of the performance of the manufacturing system, such as the minimization of the makespan or the balance of works between coworkers. The second type is specific to the agent such as the maximization of its satisfaction (regarding its preferences).
- **Knowledge base:** the agent has knowledge, which he/she uses in the manufacturing system. Among this knowledge, we can find the decision logic to select the set of machines on which the agent prefers to work. The agent has also knowledge about which assignment strategy should be selected so to achieve its goal.
- **Capabilities:** as mentioned in Franklin and Graesser (1996), the agent capabilities enable it to react rationally towards achieving a particular goal. Among these capabilities, we find the agent skills, which enable him/her, for example, to determine to which machine he/she can be assigned. The skill matrix can also contain the durations that may be needed by the agent on each machine. Based on whether the worker is novice or expert, these durations differ from one agent to another. In order to achieve its goals, the agent has physical capabilities such as the ability to recover after performing a job. For that, we consider the speed of recovery of the agent after performing a task, which varies according to the physical characteristics of the agent such as its age and gender.
- **Communication:** it represents his/her social network at work, (his/her acquaintance). The acquaintances can be the workers, which can collaborate with the agent. They can also include the supervisors. Thus, many types of communication may need to be established between agents. For instance, agents may

negotiate to switch tasks if someone is exhausted. In order to establish communication, agents need a communication protocols such as the Contract Net Protocol (Sabar et al., 2009).

3.4 System Dynamics component

The SD component describes the evolution of the agent's fatigue according to his/her behavior during the work. One of the objectives of the SD model is to describe the variation, over time, of the state variables associated with the behavioral aspects of the agent. In this study, we focus, in particular, on the evolution of the agent's fatigue.

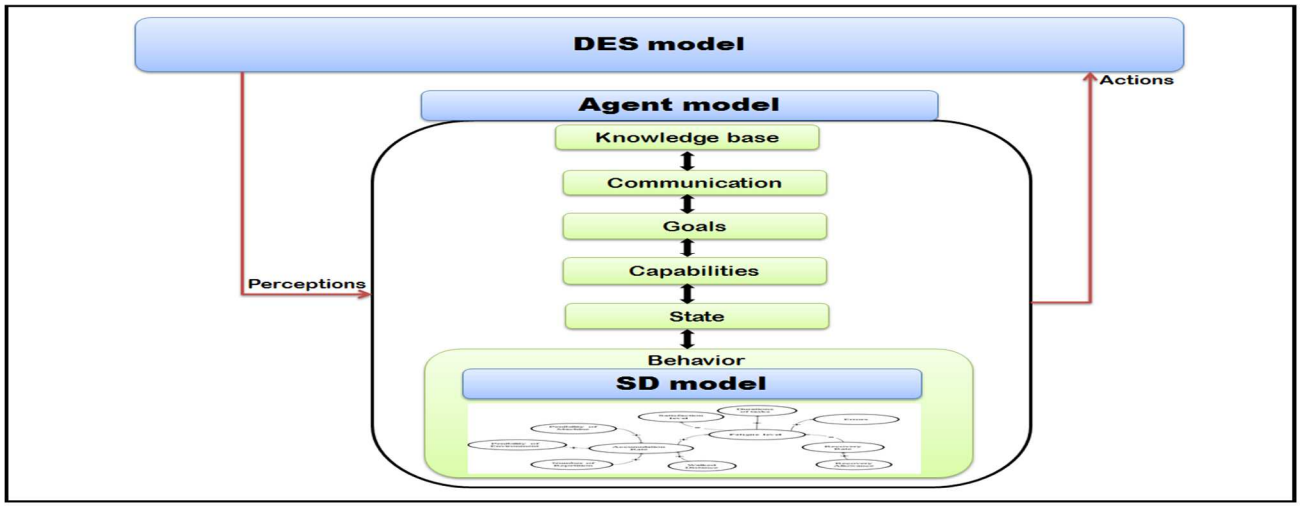


Figure 1. Overview of the proposed framework.

- **State:** in our model, the agent has different states: it can be idle, working, walking, resting and absent. State variables also describe the physical and psychological state of the agent such as Fatigue level and Satisfaction level, which indicate, respectively, the current level of fatigue and satisfaction of the worker.
- **Perception:** since we are in a simulation context, his/her perception of the environment (here the manufacturing systems and the other agents) is characterized: by the system state as it evolves in the DES component, and by what can be known about the other agents. For instance, the agent percepts the available machines in the DES model through the state variable Machine State, which is associated to each machine.
- **Behavior:** it describes the way how the state of an agent can change and how it can change the manufacturing system state. The behavior of the agent entails making some actions. For example, each agent can have assignment strategies to select to which machine he/she has to be assigned in order to satisfy his/her goal (e.g. to minimize the makespan). We can use several assignment rules, which are widely used in the literature, such as select the machine with Shortest Processing Time (SPT).

As mentioned in the related research, the behavior of workers in manufacturing is characterized by some behavioral aspects such as fatigue and satisfaction. This has an impact on how the behavioral aspects evolve along time. Let us take for example the fatigue of the agent: it increases during the period in which the agent is walking; however, not as much as when the agent is performing a tiring job.

Variations in the level of fatigue depend on several factors, such as the penibility associated with the use of machines. At the same time, these variations have a great impact on other state variables. For instance, the increase of Fatigue level leads to a decrease in Satisfaction level. As consequence, it is important to determine the causal relationships that may exist between the agent's fatigue, its factors and impacts. For that, we use a causal diagram. For illustration purposes, let us take the example of the causal diagram presented in Figure 2. This causal diagram is composed of a set of nodes, which represent the relationship between the relevant state variables used in the model.

The factors that induce fatigue are linked to the variable Fatigue Level by positive relations, since an increase of one factor causes an increase in the level of fatigue. For example, the worker's fatigue can increase with the distance he/she has walked until time t. For that, a positive relation connects the variable Walked Distance and Fatigue level. Repetitive activities also have an impact on fatigue. Therefore, there is a positive relationship connecting Fatigue level with the variable Number of Repetition, which indicates how many times the agent has repeated the job since the beginning of the simulation until time t.

The agent's fatigue has adverse effects on both the agent and DES model. These effects are represented by negative relations from the state variable Fatigue level.

On the one hand, fatigue declines some agent capabilities. For instance, the increase of fatigue leads, according to Ferjani et al. (2015 and 2017), to an increase of the task durations. Based on the human behavioral model proposed by Elkosantini and Gien (2007), fatigue causes also a decrease in the worker satisfaction. Thus, an increase of the Fatigue level leads to a decrease in the Satisfaction level. On the other hand, fatigue causes errors so that deteriorates the quality of produced parts. This means that the number of defective parts in the DES model increases.

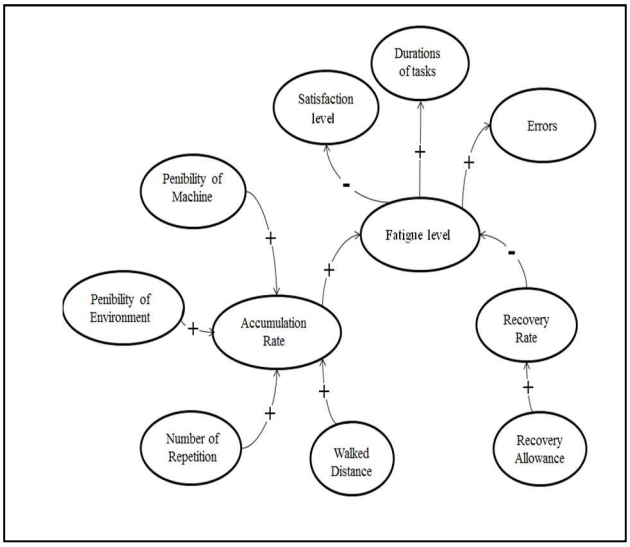


Figure 2. Causal diagram of the worker's fatigue.

As presented in Figure 2 and, according to Kahya (2007), the factors, which induce fatigue, control the speed of its accumulation. On the contrary, the recovery factors control the speed of the recovery. In our model, the parameter Recovery Allowance, which describes the physical capabilities of the agent, is used to control the speed of recovery. The variables Accumulation _Rate and Recovery_Rate control, respectively, the speed of accumulation and recovery of fatigue.

The variables in the nodes of the causal diagram (Figure 2) can be classified in two types: the first one is the discrete variables such as number of repetition of a task the worker has performed (Number of Repetitions). These variables change their value when events occur. The second type is the continuous variables such as Fatigue level and Satisfaction level. Their values change continuously, even when the agent is in the same state. Therefore, we use differential equations in order to describe the variations of the continuous variables over time. Since the way how fatigue evolves varies from one state to another, the differential equations, which are used to describe this evolution, vary also. Let us take, for example, the case where the agent is working. At this time, its fatigue grows in accordance with the penibility associated with the use of the machine (Penibility of Machine), the number of repetition of the task (Number of Repetition) and the penibility of the production environment (Penibility of Environment). In such case and based on the fatigue indicator proposed by Konz (1998), which is widely used in the ergonomic literature, the variation in the level of fatigue, when the agent is working, would be as follows:

$$dF(t)/dt = A(e,p,rep(t))(1 - F(t)) \quad (1)$$

Where *F(t)* is the current level of fatigue until time t. e and p represent, respectively, the coefficient Penibility of Environment and Penbility of Machine in the causal diagram of Figure 2. *rep(t)* represents the number of repetition of the same task.

In the case where the agent is not working, its Fatigue level decreases. We assume that the recovery depends only on the parameter Recovery Allowance. Thus, the variation in the level of fatigue, during the rest period, can be as follows:

$$dF(t)/dt = -R(rec)F(t) \quad (2)$$

Thus, the causal diagram, in Figure 2, can be translated into a flow-stock diagram. SD models are based mainly on stock (state) and flow (rate) variables. In our model, the stock variables represent the continuous state variables. For instance, the variable Fatigue level is considered as a stock variable since it describes the accumulation of fatigue during the work.

4 Implementation with ANYLOGIC

A first prototype simulation model with Anylogic has been developed to illustrate the proposed framework. The Anylogic simulation software, which is initially designed to support multiple modeling methods and their combinations (Borshchev, 2013), allows to use a combined simulation approach with several worldviews.

The different elements, which compose the DES component in our framework, can be represented using the Process Modeling Library of Anylogic, since this library supports the discrete-event modeling paradigms. The workers in the DES component are modeled as agents. For that, we represent them by a class Java since the Anylogic software is based on Java as a programming language.

Regarding the SD component, it is represented using the System Dynamics Palette of Anylogic. The state variable Fatigue level is represented by a stock, which is subjected to the variations in the flows (i.e. Accumulation Rate and Recovery Rate).

Our prototype is used to evaluate certain impacts of a given production scenarios on system performance for example, the flowtime of jobs in the system.

The evolution of the workers' fatigue along time can be described through curves, such as those presented in

Figure 3. In Figure 3, we have four workers, so that each curve corresponds to the evolution of the fatigue of each one during 120min of work. It can be noted that, during the period highlighted in Figure 3, the workers 1 and 2 are more tired than the workers 3 and 4. It can also be seen that the worker 3 has been solicited more times than the worker 1 and 2. Even so, that worker is not tired as much as the workers 1 and 2.

Acknowledgements

The authors are grateful to thank Dr. Vladimir Koltchanov, from Anylogic EUROPE, for his kind help. This project was partially supported by the CMCU-UTIQUE program 14G 1412 and by SIGMA-Clermont. Their support is greatly appreciated.

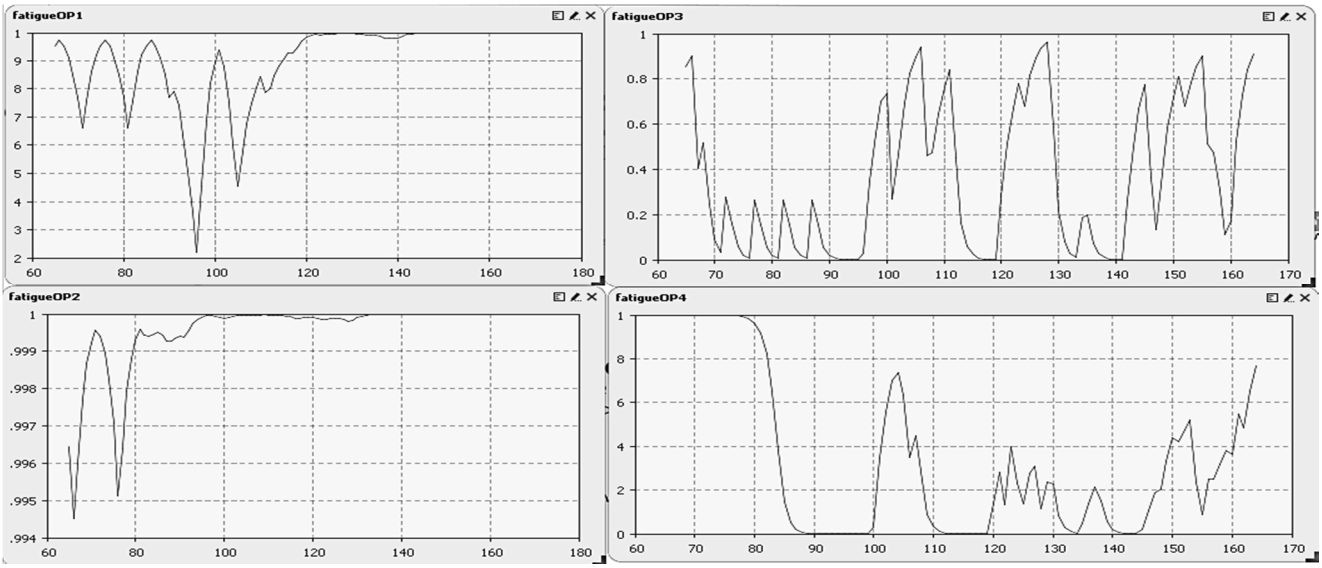


Figure 3. The evolution of workers' fatigue along time.

5 Conclusions

We have proposed a simulation modeling framework to evaluate manufacturing systems using conventional performance criteria, but also the workers' fatigue. The suggested framework combines three different paradigms (DES, SD and IA) in the same simulation model. DES is used to describe the different elements of the manufacturing system, the dynamic behavior of the system and to collect the resulting performance measures. The workers' behavior is taken into account through the use of IA concepts. The evolution of fatigue is described using SD concepts.

We have illustrated this combined approach using the ANYLOGIC simulation software. Simulation experiments show how the fatigue of each operator can evolve during his/her work depending on different organizational scenarios of the manufacturing systems. For instance, we can evaluate the impacts of different assignment strategies on workers' fatigue or of different layout of the facilities.

In order to determine the differential equations, which describe the evolution of fatigue in each state, we are based here on an indicator of fatigue (Konz, 1998), which is widely used in the literature. However, the proposed equations can be adapted with other indicators of fatigue without changing the key principles of our framework.

References

A. Borshchev. *The big book of simulation modeling: multimethod modeling with AnyLogic 6*, AnyLogic North America, 2013.

A. Ferjani, A. Ammar, H. Pierreval, and A. Trabelsi. A Heuristic approach taking operators' fatigue into account for the dynamic assignment of workforce to reduce the mean flowtime, *In International Conference on Computers and Industrial Engineering, CIE45*, 2015.

A. Ferjani, A. Ammar, H. Pierreval, and S. Elkosantini. A simulation-optimization based heuristic for the online assignment of multi-skilled workers subjected to fatigue in manufacturing systems, *Comput. Ind. Eng*, 112: 663-674, 2017.

B. Walters, J. French, and J. Barnes. Michael. Modeling the effects of crew size and crew fatigue on the control of tactical unmanned aerial vehicles (TUAVs), *In Simulation Conference, Proceedings*, 2000.

D. Dessouky and C. A. Roberts. A Review and Classification of Combined Simulation, *Comput. Ind. Eng*., 32 (2): 251– 264, 1997.

E. Kahya. The effects of job characteristics and working conditions on job performance, *Int. J. Ind. Ergon*., 37 (6): 515–523, 2007.

G. Grandjean. Fatigue in industry, *Br. J. Ind. Med.*, 36 (3): 175–186, 1979.

J. Boudreau, W. Hopp, J. O. McClain, and L. Joseph Thomas. On the interface between operations and human resources

management, *Manuf. Serv. Oper. Manag.*, 5 (3), 179–202, 2003.

J. Perez, M. P. De Looze, T. Bosch, and W. P. Neumann. Discrete event simulation as an ergonomic tool to predict workload exposures during systems design, *Int. J. Ind. Ergon.*, 44 (2): 298–306, 2014.

J. W. Forrester. Industrial dynamics: a major breakthrough for decision makers, *Harv. Bus. Rev.*, 36 (4): 37–66, 1958.

L. Padgham and P. Lambrix. Agent Capabilities: Extending BDI Theory, *In Proc. Seventeenth Natl. Conf. Artif. Intell. - AAAI*, 2000.

M. Sabar, B. Montreuil, and J.-M. Frayret. A multi-agentbased approach for personnel scheduling in assembly centers, *Eng. Appl. Artif. Intell*, 22 (7): 1080–1088, 2009.

P. Dode, M. Greig, S. Zolfaghari, and W. P. Neumann. Integrating Human Factors into Discrete Event Simulation: A proactive approach to simultaneously design for system performance and employees' well being, *Int. J. Prod. Res.*, 54 (10): 3105–3117, 2015.

S. Digiesi, G. Mossa, G. Mummolo, and P. Bari. Performance measurement and 'Personnel-Oriented' simulation of an assembly line, In *Proceedings of AMS*, 2006.

S. Digiesi, A. a. a. Kock, G. Mummolo, and J. E. Rooda. The effect of dynamic worker behavior on flow line performance, *Int. J. Prod. Econ.*, 120 (2): 368–377, 2009.

S. Elkosantini and D. Gien. Human Behavior and Social Network Simulation : Fuzzy Sets / Logic And AgentsBased Approach, *Soc. Networks*, 1: 102–109, 2007.

S. Elkosantini and D. Gien. Integration of human behavioural aspects in a dynamic model for a manufacturing system, *Int. J. Prod. Res.*, 47 (10): 2601–2623, 2009.

S. Franklin and A. Graesser. Is it an Agent, or just a program? : a taxonomy for autonomous agents, *In Intelligent agents III agent theories, architectures, and languages, Springer*, 1996.

S. Konz. Work/rest: Part II - The scientific basis (knowledge base) for the guide, *Int. J. Ind. Ergon.*, 22 (1–2): 73–99, 1998.

T. Parsons. *Politics and Social Structures*, Free Press. New York, 1979.

W. P. Neumann, W. Jörgen, R. M. L. Medbo, and S. E. Mathiassen. Production system design elements influencing productivity and ergonomics: A case study of parallel and serial flow strategies, *Int. J. Oper. Prod. Manag*, 26 (8): 904–923, 2006.

Methodology and Information Technology of Cyber-Physical-Socio Systems Integrated Modelling and Simulation

Boris Sokolov[1,2] Mikhail B. Ignatyev[2] Karim Benyamna[3] Dmitri Ivanov[4] Ekaterina Rostova[1]

[1]St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Russia
sokolov_boris@inbox.ru, rostovae@mail.ru
[2]St. Petersburg State University of Aerospace Instrumentation (SUAI), Russia
ignatmb@mail.ru
[3]St. Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO Univ.), Russia
benyamna.karim@gmail.com
[4] Department of Business Administration, Berlin School of Economics and Law, Germany
dmitri.ivanov@mail.ru

Abstract

The main objects of our investigation are cyber-physical-socio space and systems (CPSS). The CPSS is the fusion of the physical space, the cyber space, and the social space. The problem of CPSS integrated modelling and simulation is an actual modern problem. The solution of this problem involves interdisciplinary research by specialists in mathematics, economics, biology, physics, and computer technologies. Therefore, the paper presents the results of research in the field of CPSS modelling and multi-agent simulation.

Keywords: interdisplinary research, cyber-physical-socio space and systems, cybernetics, control and management, modelling and multi-agent simulation

1 Introduction

Today we come to the realization that a transformation from an industrial society to an informational society should be guided. Now we live in cyber-physical-socio space and systems (CPSS) (Lee, 2008). It is also called cyber-physical society (Zhuge, 2010). CPSS in contrast with cyber-physical systems (CPS) consist of not only cyberspace and physical space, but also human knowledge, mental capabilities, and sociocultural elements.

The technologies for control and management of transformation from an industrial society to an informational society need regulation and structuring at a macro and micro level. This inspires a renewed interest in the theoretical background of control and management problems. Unfortunately, logically relevant chain of fundamental notions: Cybernetics – Control – Informational processes – Universal transformer of the information (computer, cybernetic machine) was split. An expansion of computer technologies caused an illusion of their ability to solve any problem. The imperfection of these technologies has already caused catastrophes that let American and European scientists proclaim establishing "Risk society" rather than "Informational" one.

Two main reasons stimulate importance of the new cybernetics in the modern world (Bir, 1963; Okhtilev et al, 2006). The first one deals with a problem of complexity which has various applications and aspects (structural complexity, complexity of functioning, complexity of decision making, etc.). The second reason is the lack of holistic (system) thinking in the IT industry. The problem of complexity control and management involve interdisciplinary research by specialists in mathematics, economics, sociology, biology, physics, and computer technologies. However, the founders of cybernetics viewed its laws as much more universal than they are really considered in today's social and business systems. Therefore, the paper presents the results of interdisciplinary research in the field of computer modeling and decision support systems in social-cybernetics objects. This field is called neocybernetics (von Foerster, 1987; Okhtilev, 2006).

2 Interdisciplinary Branch of System Knowledge as the Basis for Cyber-Physical-Socio Systems Integrated Modelling and Simulation

As a result of the continuing scientific and technical revolution which beginning is dated the middle of the XX century, there appeared a significant number of complex objects (nuclear power plants, space equipment, electronics, computers, etc.) research, the description, design and management of which presents essential difficulties and problems. CPSS are striking examples of complex systems. Figure 1 illustrates the evolution from the cyberspace and systems to the cyber-physical-social space and systems (Lee, 2008). In Figure 1 part (a) depicts two types of cyber space: the first one only allows users to read the information in the cyber space like the Web, and the other one allows users to read and write

information in the cyber space like the Web 2.0. Both rely on humans to add information to the cyber space in order to share it with others. Part (b) in Figure 1 depicts the extension of the cyber space to the physical space through various sensors. Any significant information in the physical space can be automatically sensed, stored and transmitted through the cyber space. Web of things can be considered as a kind of cyber-physical space. Part (c) in Figure 1 depicts an important extension of part (b): user behaviors can be sensed and feedbacked to the cyber space for analyzing the patterns of behaviors, and humans can remotely control the actuators to behave in the physical space through the cyber space. This enables the cyber space to adapt its services according to the feedback since behavior change may indicate some psychological change (Zhuge, 2010). Part (d) in Figure 1 depicts a simple cyber-physical-social space. Not only individual's behaviors, but also social interactions can be feedback into the cyberspace for further processing. Users are considered with their social characteristics and relations rather than as isolated individuals. Sensors are limited in their ability to collect all information in the physical space, so users still need to directly collect the significant information in the physical space and then put it into the cyberspace after analysis (including experiment). Users can also manipulate physical objects in the physical space, which can also be feedbacked into the cyber space to reflect the real-time situation (Zhuge, 2011). Users' status, interests and knowledge evolve with social interaction and operations in the cyber space.

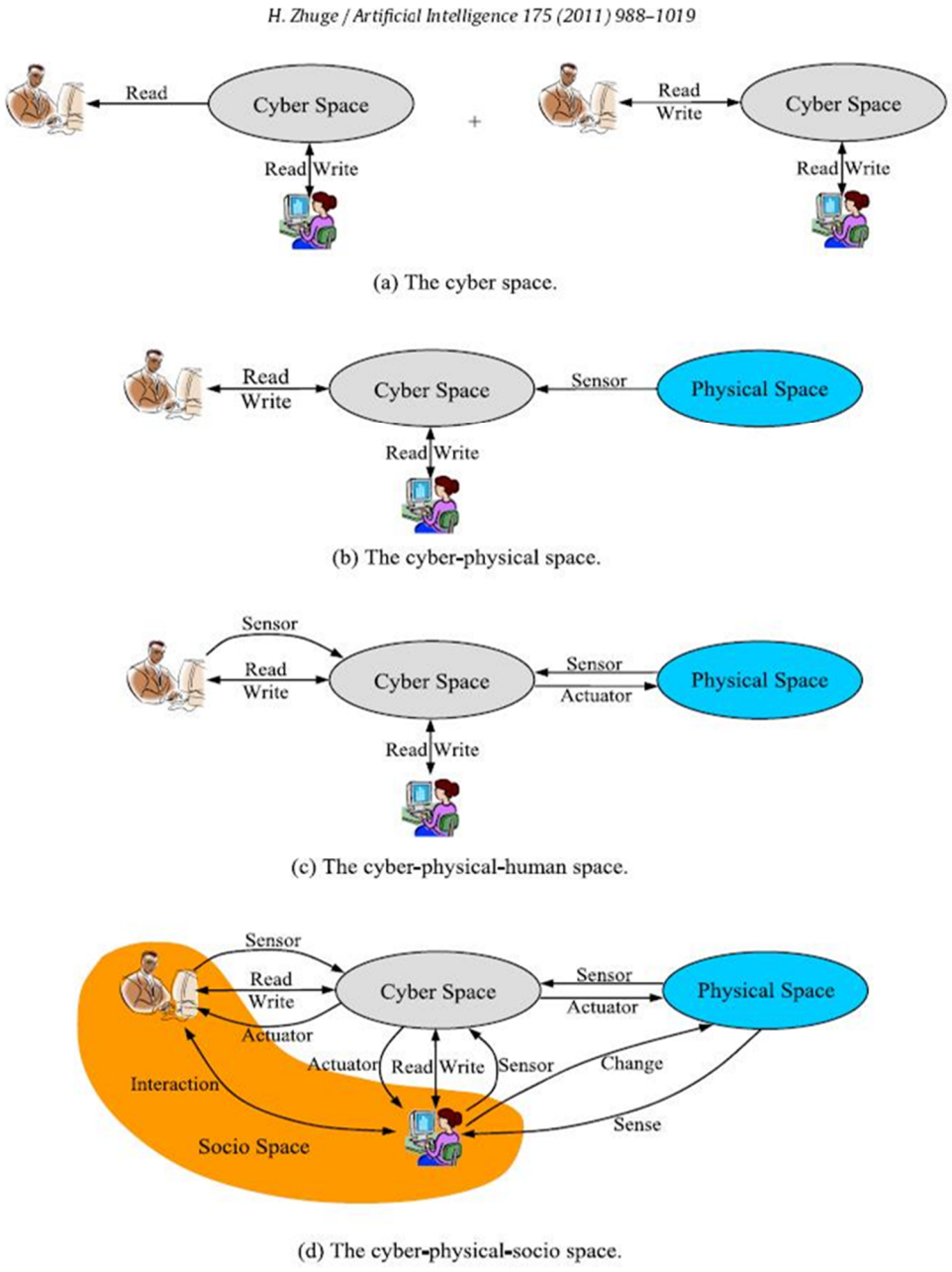


Figure 1. The emerging cyber-physical-socio space.

So today, we come to the realization of a transformation from an industrial society which is based on cyber-physical space to an informational society which is interrelated with cyber-physical-socio space and systems. The technologies for control of this transformation need regulation and structuring at a macro and micro level. This inspires a renewed interest in the theoretical and practical background of complexity control and management problems. There exist different aspects of CPSS. For example, CPSS are characterized by following properties: high complexity and dimensionality with such features as redundancy, multifunctionality, distributed elements, unification, uniformity of main elements, subsystems and interrelations; structure dynamics; nonlinear uncertain behavior; hierarchical network structure; nonequilibrium; uncertainty of observer selection and interaction with him; dynamic rules and regulations; counteracting and amplifying relationships with self-excitation; possible chaotic behavior; no element has enough information about the whole system; selective sensitivity to input actions (dynamic robustness and adaptation); the response time is greater than the time between input actions, and it is greater than the time of supervention; the real object of control cannot have a complete and reliable description (in accordance with Bremerman's limit and Godel's theorem) (Okhtilev et al, 2006; Ignatyev, 2008). The specified objective circumstances which are connected with CPSS have resulted in need at the beginning of the XXI century of formation in a wide range of experts of system outlook and the relevant system branch of scientific knowledge. It has been shown in the works (Lee, 2008; Ignatyev, 2008) that speaking about the interdisciplinary branch of the system's knowledge, it is expedient to select in it two big sections (block) — the block of fundamental system info-cybernetic knowledge and the block of applied system info-cybernetic knowledge. In the first of the listed blocks, the defining role is played by three scientific directions — the general theory of systems, cybernetics and informatics.

As for the general theory of systems (systemology), this scientific direction sets the task to construct the general scientific basis for systems of any nature. The central concepts of the general theory of systems are the concepts of an open system, i.e. the system interacting with the environment surrounding it, both complex and big. The mathematical basis for the general theory of systems can reasonably enough be considered as a certain interpretation of the basis for mathematics, mainly theories of the relations (the concept of the relation is fundamental both in mathematics, and in system researches), theories of mathematical structures and the theory of categories and functors. However, except for a number of recognized general provisions in general for the present, there is no uniform understanding of in what way this theory has to be applied.

As regards informatics, this scientific direction is connected with the development of methods and means of collecting, storage, transfer, representation, processing and information security. Speaking about the processes of interaction of cybernetics with informatics it should be noted, first, that historically the last one considerably developed into a subsoil of traditional cybernetics, actually on uniform technical base — computer facilities and means of communication and data transmission, and, secondly, cybernetics, being a science about the general laws and regularities of management and communication, has objectively been forced in recent years to deal with the second round of rapprochement of cybernetics and informatics. There is an active terminological and substantial interpenetration of these scientific directions, issues of use of information for the benefit of management. Therefore, the methods, technologies and means developed into an informatics that actively take root into cybernetics within such new scientific directions as: information management, different types of intellectual management (situational, neuromanagement, the management based on knowledge, on the basis of evolutionary algorithms, multi-agent management and etc.). These types of intellectual management are based, in turn, on the appropriate intellectual information technologies (IIT) focused on symbolic information processing.

Cybernetics is the general theory of management. Initially with the founder of cybernetics N. Winer in 1948 in his book "Cybernetics or Management and Communication in an Animal and the Car", it was emphasized that this science is about management, communication and processing of information in systems of any nature (Wiener, 1950). At the same time, the main goal of the research conducted within the specified science consisted in identification and establishment of the most general laws of functioning to which submit as the operated objects, and the corresponding managing directors of a subsystem without regard to their nature. The classical cybernetics has reduced all earlier existing views of management processes into uniform systems and has proved its completeness and generality. In other words, it has shown in detail the raised power of system approach to the solution of complex problems (Ignatyev, 2008). The most developed direction in cybernetics was the theory of management of dynamic technical systems within which numerous outstanding fundamental and applied scientific results have been received by multiple experts (Heikki, 2006; Okhtilev et al, 2006; Ignatyev, 2008; Zhuge, 2010, 2011).

In turn, cybernetic terminology gets into informatics and computer facilities. Today, in particular, concepts and, respectively, strategies of adaptive and proactive computer systems, adaptive management and the adaptive enterprise are very popular in the IT industry. Those strategies are intensively developed by the companies IBM, Intel Research, Hewlett-Packard, Microsoft, Sun, etc. (Okhtilev et al, 2006). At the same time, the material basis for the realization of technologies of the operated self-organization is created. In the modern business systems (BS) only those organizations obtain success in which development of IT architecture is focused on the Web services and technologies allowing to effectively decentralize traditional systems of decision-making, turning them into self-regulating subsystems. Interaction of cybernetics (neocybernetics) and informatics with the

general theory of systems is carried out in several directions. The first of these directions is directly connected with the generalized description of objects and subjects of management on the basis of the new formalistic approaches developed in a modern systemology to which it is possible to refer, for example, structural and mathematical and category-functor's approaches (Okhtilev et al, 2006). In this regard, it is also possible to note interesting scientific results which have been received in a qualimetry of models and poly-model complexes and can be used in informatics and cybernetics. The methods and algorithms of decomposition (composition), aggregation (disaggregations), and coordination developed in the general theory of systems in relation to objects of any nature are widely used in cybernetics and informatics also in the solution of the problems of collecting, storage, transfer, representation, processing, information security, and also management of complex objects. On the other hand, it has been shown in the works (Wiener, 1950; Foerster, 1987; Okhtilev, 2006) that the approaches developed in the classical theory of management of technical objects and also in modern informatics can be applied successfully to the organization of processes of management of quality of models and poly-model complexes, and also at their structural and parametric adaptation.

3 Methodological and Technical Basis for CPSS Integrated Modelling and Simulation

During our investigation, we describe the main classes of CPSS integrated modeling tasks. For these aims, we use structure-dynamics control (SDC) theory. The methodological basis of this theory includes the methodologies of generalized system analysis and the modern optimal control theory for CTS with reconfigurable structures. The dynamic interpretation of SDC processes lets apply the results, previously received in the theory of dynamic system's stability, ability, failure tolerance, effectiveness and sensitivity, for CPSS analysis problems. The existence of various alternative descriptions for CPSS elements and control subsystems gives an opportunity of adaptive model selection (synthesis) for program control in changing environment. Therefore, we considered two general actual problems of the CPSS structure-dynamics investigation. Those are the problem of selection of optimal CPSS structure-dynamics control programs at different states of the environment and the problem of parametric and structural adaptation of models describing CPSS structure-dynamics control. In this case the adaptive control should include the following main phases: parametric and structural adaptation of SDC models and algorithms to previous and current states of objects-in-service (SO), of control subsystems (CS), and of the environment; integrated scheduling of CPSS operation (construction of SDC programs); simulation of CPSS operation, according to the schedules, for different variants of control decisions in real situations; structural and parametric adaptation of the schedule, control inputs, models, algorithms, and CPSS programs to possible (predicted by simulation) states of SO, CS, and the environment.

During our investigations, the main phases and steps of a program construction procedure for optimal structure-dynamics control in CPSS were proposed (Okhtilev et al, 2006). At the first phase forming (generation) of allowable multi-structural macro-states is being performed. In other words, a structure-functional synthesis of a new CPSS make-up should be performed in accordance with an actual or forecasted situation. Here the first-phase problems come to CPSS structure-functional synthesis. In the second phase, a single multi-structural macro-state is being selected, and adaptive plans (programs) of the CPSS transition to the selected macro-state are being constructed. These plans should specify transition programs, as well as programs of stable CPSS operation in intermediate multi-structural macro-states. The second phase of program construction is aimed at a solution of multi-level multi-stage optimization problems.

One of the main opportunities of the proposed method of CPSS SDC program construction is that besides the vector of program control, we receive a preferable multi-structural macro-state of the CPSS at the end point. This is the state of CPSS reliable operation in the current (forecasted) situation. The combined methods and algorithms of optimal program construction for structure-dynamics control in centralized and non-centralized modes of CPSS operation were developed (Okhtilev et al, 2006). Classification and analysis of perturbation factors having an influence upon the operation of CPSS were performed. Variants of perturbation-factors descriptions were considered in CPSS SDC models. In our opinion, an integrated simulation of uncertainty factors with all adequate models and forms of description should be used during the investigation of CPSS SDC. Moreover, the abilities of CPSS management should be estimated in both normal mode of operation and emergency situations. It is important to estimate the destruction "abilities" of perturbation impacts. In this case, the investigation of CPSS functioning should include the following phases: determining of scenarios for CPSS environment, particularly determining of extreme situations and impacts that can have catastrophic results; analysis of CPSS operation in a normal mode on the basis of a priori probability information (if any), simulation, and processing of expert information through the theory of subjective probabilities and theory of fuzzy sets; repetition of item b for the main extreme situations and estimation of guaranteed results of CPSS operation in these situations; computing of general (integral) efficiency measures of CPSS structure-dynamics control. Algorithms of parametric and structural adaptation of CPSS SDC models were proposed (Okhtilev et al, 2006). The algorithms will be based on the methods of fuzzy clusterization, on the methods of hierarchy analysis, on biological adaptation mechanisms, and on the methods of a joint use of analytical and simulation models. We illustrate our methodology by two examples.

4 Example of Lingvo-Combinatorial Modelling

We often use a natural language to describe systems. We propose to transfer this natural language description to mathematical equations.

For example, we have a sentence:

$$WORD1 + WORD2 + WORD3 \quad (1)$$

Here we assign words and only imply the meaning of words, the meaning (sense) is only implied but not designated.

We propose to assign meaning in the following form:

$$(WORD1) * (SENSE1) + (WORD2) * (SENSE2) + (WORD3) * (SENSE3) = 0 \quad (2)$$

This equation (2) can be represented in the following form:

$$A1 * E1 + A2 * E2 + A3 * E3 = 0 \quad (3)$$

Where Ai, i = 1, 2, 3, will denote words from English Appearance and Ei will denote senses from English Essence. The equations (2) and (3) are the model of the sentence (1). When we have a mathematical equation in the form F(x1, x2, x3) = 0, we can turn such a form by means of differentiation where the partial derivatives are the appearances and the derivatives with respect to time are the essences. This model is an algebraic ring and we can resolve this equation with respect to the appearances Ai or the essences Ei (Ignatyev, 2008):

$$\begin{aligned} A1 &= U1*E2 - U2*E3 \\ A2 &= - U1*E1 + U3*E3 \\ A3 &= - U2*E1 - U3*E2 \end{aligned} \quad (4)$$

or

$$\begin{aligned} A1 &= U1*A2 - U2*A3 \\ A2 &= - U1*A1 + U3*A3 \\ A3 &= -U2*A1 - U3*A2 \end{aligned} \quad (5)$$

Where U1, U2, U3 are arbitrary coefficients, can be used for solution of different tasks on the initial manifold (2) or (3). In general, if we have n variables in our system and m manifolds, constraints, then the number of arbitrary coefficients S will be defined as the number of combinations from n to m+1 (Ignatyev, 2008), as shown in Table 1.

Table 1. Number of combinations from N to M+1.

n /m	1	2	3	4	5	6	7	8
2	1							
3	3	1						
4	6	4	1					
5	10	10	5	1				
6	15	20	15	6	1			
7	21	35	35	21	7	1		
8	28	56	70	56	28	8	1	
9	36	84	126	126	84	36	9	1

$$S = C_n^{m+1} \quad n > m \quad (6)$$

The formula (6) is the basic law of cybernetics, informatics and synergetics for complex systems (Ignatyev, 2008). The number of arbitrary coefficients is the measure of uncertainty. Usually, when solving mathematical systems, the number of variables is equal to the number of equations. In practice, we frequently do not know how many constraints, there are on our variables. Combinatorial simulation makes it possible to simulate and study the systems with uncertainty on the basis of incomplete information. The problem of simulation of condition, guaranteeing the existence of maximum adaptability, is considered.

It is supposed that the behavior of a system with n variables is given with an accuracy of m intersecting manifolds, n > m. If the system is considered as a multidimensional generator where at least a part of the variables interacts with environment variables, and if the objective of the system is to decrease the functional of discoordination between them (1… k), the system control unit has two instruments of impact, a and b, upon the system. First, this is the tuning – the changing of uncertain coefficients in the structure of the differential equations of the system, taking into account that the greater number of these coefficients implies a more accurate system response to a changing environment. Second, this is the learning – the imposing new restrictions on the system behavior. The number of arbitrary coefficients, in the structure of equivalent equations, changes in the process of learning, of consecutive imposing new and new restrictions on the system behavior. In the systems with more than six variables the number of arbitrary coefficients increases first, and then, passing through the maximum begins to decrease. This phenomenon makes it possible to explain the processes of system growth, complication and death. The existence of maximum adaptability phenomenon is observed in numerous biological, economical and physical-engineering systems. It is important that we describe a system with a full sum of combinations and

have all the variants of decisions. The linguo-combinatorial simulation is a useful heuristic approach for investigation of complex, poorly formalized systems. Natural language is the main intellectual product of mankind; the structure of natural language reflects the structure of natural intellect of mankind and its certain representatives on the level of consciousness and unconsciousness. Linguo-combinatorial simulation is the calculation, which allows to extract the senses from texts. L. Wittgenstein wanted to have the calculation of senses. In our calculation we have the three groups of variables: the first group – the words of natural language Ai, the second group — the essences Ei, which can be the internal language of brain (Ignatyev, 2008); we can have the different natural languages, but we have only one internal language of brain; this hypothesis opens a new way for experimental investigation; the third group of variables – the arbitrary coefficients, structural uncertainty in our model, which we can use for adaptation in translation processes and etc.

5 Example of Transport Networks Simulation

As an example of the integrated modeling and simulation based on the CPSS, we developed a simulation program (Figure 2) and an experimental stand (Figure 3) with a transport network (for example, railroad) and some production and warehouse facilities are currently under development. The railroad is provided with multiple sensors, for example, the RFID, which provides information about the position and the speed of bypassing locomotives.

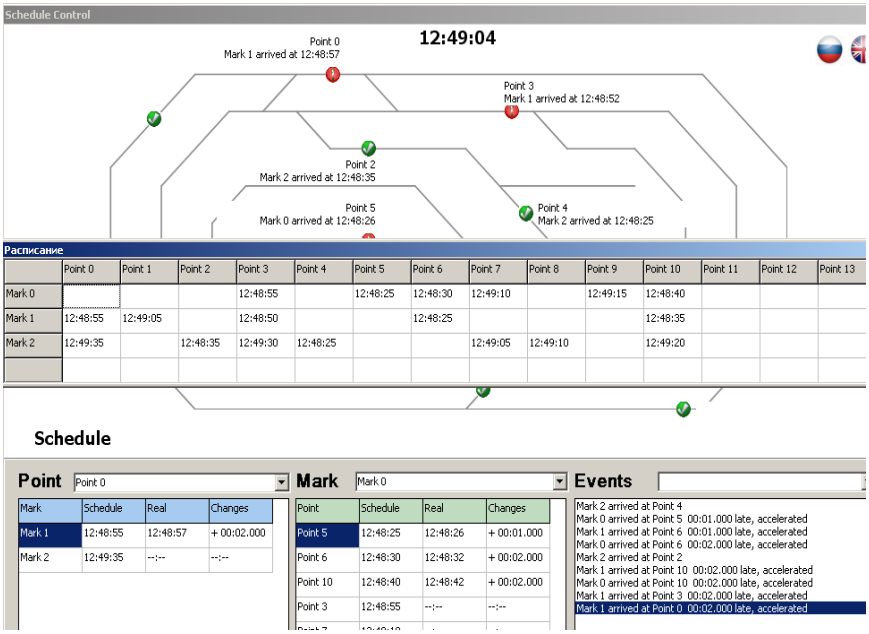


Figure 2. Screenshot from the transport network simulation program.

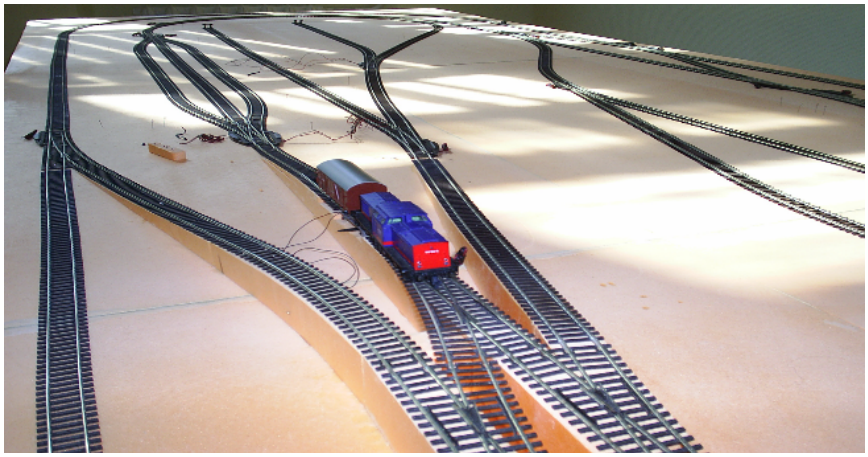


Figure 3. The experimental stand.

We note that the RFID experimental environment is not intended (at least, in its current version) for a full implementation of the developed models. It is much simpler than the modeling framework and serves to gather experimental data for the modeling complex. The modeling complex itself is implemented in a special software environment, which contains a simulation and optimization engine of CS planning and execution control, a Web platform, an ERP system, and APS system, and a SCEM system. The kernel of the computational framework is the decision modeling component, that is, the simulation and optimization engine. The schedule optimization is based on an optimal control algorithm that is launched by a heuristic solution, the so-called first approach. The seeking for the optimality and the CS scheduling level is enhanced by simultaneous optimizing and balancing interrelated CS functional, organizational and information structures. The schedules can be analyzed with regard to performance indicators and different execution scenarios with different perturbations. Subsequently, parameters of the CS structures and the environment can be tuned if the decision-maker is not satisfied with the values of performance indicators. In analyzing the impact of the scale and location of the adaptation steps on the CS performance, it becomes possible to justify methodically the requirements for the RFID functionalities, the stages of a CS for the RFID element locations, and the processing information from RFID. In particular, possible discrepancies between actual needs for a wireless solution of CS control problems and the total costs of ownership regarding RFID can be analyzed. In addition, processing information from RFID can be subordinated to different management and operation decision-making levels (according to the developed multi-loop adaptation framework). Pilot RFID devices with reconfigurable functional structure are developed (Kurve et al, 2013). In order to simulate the RFID-based transport network, we have created a prototype of a simulation model that reproduces a real railway network. To do this, we could use different approaches, but the most suitable are the multi-agent system (MAS) modelling (Niazi et al, 2011; Kurve et al, 2013). To realize the simulation model, we could use different tools made for modeling a MAS (Okhtilev et al, 2006; Lee, 2008; Niazi et al, 2011; Kurve et al, 2013) but most of them have a restrictive or a paid licensing policy. So we managed to develop from scratch a prototype of a simulation environment based on the C++ programming language and the MAS approaches. The first step in creating a simulation model is to define the time framework (Niazi et al, 2011) implemented in the form of a main event loop. In the loop, we have two functions. The first may or may not create a random disturbance in the locomotive speed and the maximum speed. The second function simulates the agent behavior. This system enables us to simulate a basic railroad network and test the RFID infrastructure on a realistic model. Of course, as any system, this one has its advantages and disadvantages. As regards advantages, they include: the technologies used to create this model (C++ language), it provides us with a great flexibility in terms of functionality, allowing for modification and implementation of any kind of logic we want; agent-based modeling is a powerful method allowing a large number of enhancements in the behavior of the system. Additionally, it enables us to define a logic of each individual locomotive, which is close to how decisions are made in a real system; define the system behavior by an independent entity allows greater scalability, as the complexity of the system is linear to the number of entities.

6 Conclusions

Dynamic multiple-model descriptions of CPSS functioning at different stages of their lifecycle are proposed in the paper. Different types of models (analytical-simulation, logical algebraic, logical-linguistic models) were proposed for description and study of the main attributes of the CPSS (Yusupov et al, 2011; Geida et al, 2015; Yusupov, 2009). Joint use of diverse models in the framework of poly-model systems, allows one to improve the flexibility and adaptability of IDSS, as well as to compensate the drawbacks of one class of models by the advantages of the other (Okhtilev et al, 2006; Ignatyev, 2008).

Acknowledgements

The research described in this paper is partially supported by the Russian Foundation for Basic Research (grants 15-07-08391, 15-08-08459, 16-07-00779, 16-08-00510, 16-08-01277, 16-29-09482-ofi-i, 17-08-00797, 17-06-00108, 17-01-00139, 17-20-01214, 17-29-07073-ofi-i), grant 074-U01 (ITMO University), state order of the Ministry of Education and Science of the Russian Federation №2.3135.2017/4.6, state research 0073–2014–0009, 0073–2015–0007, International project ERASMUS +, Capacity building in higher education, № 73751-EPP-1-2016-1-DE-EPPKA2-CBHE-JP,
Innovative teaching and learning strategies in open modelling and simulation environment for student-centered engineering education.

References

S. Bir. *Cybernetics and production control*. Fizmatlit. 1963.

H. von Foerster. *Cybernetics. Encyclopedia of Artificial Intelligence*. John Wiley and Sons. 1987.

A.S. Geida, I.V. Lysenko, and R.M Yusupov. Main Concepts and Principles for Information Technologies Operational Properties Research. *SPIIRAS Proceedings,* 42:5–36, 2015.

H. Heikki. Neocybernetics in Biological Systems. Helsinki University of Technology, Control Engineering Laboratory. Report 151, p. 273, August 2006.

M.B. Ignatyev. Semantics and selforganization in nanoscale physics. *International Journal of Computing Anticipatory Systems,* 22:17–23, 2008.

A. Kurve, K. Kotobi, and G. Kesidis. An agent-based framework for performance modeling of an optimistic parallel discrete event simulator. *Complex Adaptive Systems Modeling*, 1(1):12, 2013. doi: 10.1186/2194-3206-1-12

E.A. Lee. Cyber physical systems: Design challenges. *2008 11th IEEE International Symposium on Object and*

Component-Oriented Real-Time Distributed Computing (ISORC), 11:363-369, 2008. doi: 10.1109/ISORC.2008.25

M. Niazi, A. Hussain. Agent-based computing from multi-agent systems to agent-based models: a visual survey. *Scientometrics*, 89(2):479–499, 2011. doi: 10.1007/s11192-011-0468-9

M.Yu. Okhtilev, B.V. Sokolov, and R.M. Yusupov. *Intellectual technologies for monitoring and control of structural dynamics of complex technical plants*. Nauka Publishers. 2006.

N. Wiener. *The Human Use of Human Beings: Cybernetics and Society*. Da Capo Press. 1950.

R.M. Yusupov, B.V. Sokolov, A.I. Ptushkin, A.V. Ikonnikova, S.A. Potryasaev, and E.G.Tsivirko. Research problems analysis of artificial objects lifecycle management. *SPIIRAS Proceedings,* 16:37–109, 2011.

R.M. Yusupov. About the impact of information and communication technologies on the national security assurance in the environment of the information society forming. *SPIIRAS Proceedings,* 8:21–33, 2009.

H. Zhuge. Cyber physical society. *Proceedings of the 6th International Conference on Semantics, Knowledge and Grids*, 1:1-8, 2010.

H. Zhuge. Semantic linking through spaces for cyber-physical-socio intelligence: a methodology. *Artificial Intelligence,* 175:988–1019, 2011. doi: 10.1016/j.artint.2010.09.009

Reliable Detection of a Variance Increase in a Critical Process Variable

Mika Pylvänäinen Toni Liedes

Mechatronics and Machine Diagnostics Research Unit, University of Oulu, Finland,
{mika.pylvanainen,toni.liedes}@oulu.fi

Abstract

Industrial process failures can be often seen as a variance increase in a measured process variable. The objective of this research was to investigate if stochastic Autoregressive Moving Average, abbreviated as ARMA, and Generalized Autoregressive Conditionally Heteroscedastic, abbreviated as GARCH, time series modelling are feasible methods for the reliable detection of gradually increasing variance in the process variable. A case study was conducted for the reliable detection of increased pressure variance that indicates a harmful air leakage in a vacuum chamber in a paper machine. Variance in the chamber pressure was artificially gradually increased, a combined ARMA+GARCH time series model was fitted to it and the variance vector was determined. An abnormally high variance was detected from the variance vector using a specified detection limit and detection sensitivity. According to the simulation results, by controlling the variance vector extracted from the combined ARMA+GARCH time series model, a very slight variance increase in the process variable can be detected more reliably than detecting it from the moving variance vector computed directly from the process variable.

Keywords: condition monitoring, time series analysis, SPC

1 Introduction

The variance increase of the critical process variable is often an early sign of an abnormal process behavior (Jandhyala et al., 2002; Du et al., 2010). It can be also a sign of a phenomenon that may escalate to an unplanned process shutdown or severe damage in a critical process component (Rzadkowski et al., 2016). This can cause a significant loss in production capacity due to an unplanned maintenance or the long delivery time of a spare part. Therefore, it is crucial to reliably detect an abnormal variance increase in real time process condition monitoring. This enables a transition from reactive maintenance to condition-based maintenance.

In the paper manufacturing process, there are several critical process variables whose variance increase is a sign of incipient failure.

One typical example is increased gauge pressure variance, later called pressure variance, in a certain vacuum chamber in a paper machine. In a stable and faultless paper manufacturing process, the mean of the vacuum chamber pressure is constant while its variance inherently fluctuates within a certain range.

According to process specialists, air leakage in the vacuum chamber gradually increases variance in the chamber pressure while the pressure mean remains constant. When air leakage increases, also the pressure variance increases until the pressure value exceeds its tolerance limit. This triggers an operation sequence of a certain mechanical function, the sound of which is subjectively interpreted as a sign of air leakage. Instead of the subjective interpretation of a sound in a noisy paper machine environment, an incipient air leakage in the vacuum chamber could be reliably identified if an abnormal variance increase is detected from the chamber pressure. This improves product quality, enables preventive actions and helps avoid costly process breakdowns.

Numerous time series analysis methods are used in economics to model and predict econometric phenomena. Two commonly used analysis methods in economics are Autoregressive Moving Average, abbreviated as ARMA, and Generalized Autoregressive Conditionally Heteroscedastic, abbreviated as GARCH, time series modelling. Although several papers (Wang et al., 2002; Tao et al., 2010) have been published on the separate use of these methods in condition monitoring, fewer studies (Pham et al., 2010; Caesarendra et al., 2011) have been published on the joint use of the ARMA and GARCH time series models in this field. These studies prove that the combined ARMA +GARCH time series model provides a versatile and powerful toolkit for condition monitoring.

The objective of this research was to investigate if the combined ARMA+GARCH time series modelling is a feasible method for the reliable detection of increased variance in a critical process variable. Data analysis was done by R software.

A case study was conducted for the detection of increased pressure variance in a vacuum chamber in the paper machine. At the beginning of the study, chamber pressure data were recorded at 8 s sampling intervals in a 140 h period. During that time, there was no air leakage in the vacuum chamber.

Then the ARMA model was fitted to the data. Autocorrelation of the squared residuals of the fitted ARMA model indicated that they still contain information so the residuals were further modelled using the GARCH model. The squared residuals of the fitted GARCH model were no longer autocorrelated. Thus all the information included in the chamber pressure data was decomposed to the coefficients of the fitted ARMA +GARCH model. The model defines a typical behavior for vacuum pressure when air leakage does not occur.

Due to missing air leakages in the vacuum chamber during the study, the vacuum pressure vector including the air leakage effect was artificially created by first simulating the pressure vector using the combined ARMA+GARCH model with coefficients representing the process without air leakage. Then a vector with mean zero and slightly increasing variance was summed to the simulated pressure vector to mimic air leakage. The sum vector represents the chamber pressure in an incipient air leakage condition. The goal in the next steps of the study is to detect the systematic variance increase in the chamber pressure with minimum delay.

An artificially created pressure vector was modelled by means of the combined ARMA+GARCH model using the same number of coefficients as in the stable process model. The variance vector extracted from the aforementioned model illustrates the variance behavior of the pressure vector. It is challenging to detect variance increase caused by air leakage due to an inherent variance fluctuation even when the air leakage is not present. An alarm for an abnormally high variance in the vacuum pressure is triggered when the variance vector values exceed a quantile that is defined using an adjustable detection limit, abbreviated as p_k. The quantile divides an empirical distribution of past variance values so that p_k proportion of variance values are smaller than the quantile. To avoid false alarms, adjustable detection sensitivity, abbreviated as s, is used to specify the minimum number of consecutive values of the variance vector above the quantile before the alarm is triggered. Once the alarm is triggered, a delay, i.e. number of pressure observations from the start of the leakage until its detection, is recorded.

The sequence described above was simulated $27 \cdot 10^3$ rounds. According to the simulation results, it is possible to detect an almost invisible variance increase in the vacuum chamber pressure more reliably than detecting it from the moving variance vector computed from the process variable.

2 Time Series Models and Detection Parameters

In this chapter the Autoregressive model, the Moving Average model and their combination called Autoregressive Moving Average model are examined. Then the Generalized Autoregressive Conditionally Heteroscedastic model is introduced and description is given how the autocorrelated time series vector is decomposed by the combined use the Autoregressive Moving Average model and the Generalized Autoregressive Conditionally Heteroscedastic model. Finally, a detection limit and detection sensitivity of abnormally high variance detection in the time series vector is examined.

2.1 Autoregressive Moving Average Model

The ARMA model is the most popular class of linear time series models. ARMA models are commonly used to model linear dynamic structures and, to describe a linear relationship among lagged variables, i.e. when the variable is autocorrelated. The ARMA model consists of two submodels, the Autoregressive model, abbreviated as AR, and the Moving Average model, abbreviated as MA.

The Autoregressive model of order p, abbreviated as AR(p), is of the form

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + w_t, \quad (1)$$

where x_t is a stationary time series with mean zero; ϕ_1, ϕ_2, ..., ϕ_p are constants ($\phi_p \neq 0$) and w_t is a Gaussian white noise with mean zero and variance σ_w^2. The AR(p) model assumes that the current value of time series x_t is defined as a sum of the linear combination of p past time series values and the white noise w_t (Shumway et al., 2011; Box et al., 1976).

The Moving Average model of order q, abbreviated MA(q) is of the form

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + ... + \theta_p w_{t-q}, \quad (2)$$

where x_t is a stationary time series with mean zero; θ_1, θ_2, ..., θ_q ($\theta_q \neq 0$) are the constant parameters and w_t is the Gaussian white noise with mean zero and variance σ_w^2. The MA(q) model assumes that the current value of time series x_t is defined as the linear combination of the current and q previous values of the white noise w_t (Shumway et al., 2011; Box et al., 1976).

When the Autoregressive model AR(p) and the Moving Average model MA(q) are combined, the Autoregressive Moving Average model of order p and q, abbreviated as ARMA(p, q) is defined. ARMA(p, q) is of the form

$$x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + ... + \theta_p w_{t-q}. \quad (3)$$

The ARMA model assumes that the variance of the time series x_t is constant. This is not always the case. Thus varying variance cannot be modelled using the ARMA model and therefore its residuals are not Gaussian white noise, i.e. they still contain information (Shumway et al., 2011; Box et al., 1976).

2.2 Generalized Autoregressive Conditionally Heteroscedastic Model

The Generalized Autoregressive Conditionally Heteroscedastic model, abbreviated as GARCH, is frequently used in economics and finance to model the varying variance of the time series x_t, i.e. when the variance of the time series is autocorrelated. The GARCH modelling is used to model such information from time series x_t that is not possible with the ARMA modelling. For example, autocorrelated squared residuals of the ARMA model indicate that they still contain information that cannot be modelled by means of the ARMA model. Then the GARCH is used to model the information included in the residuals of the ARMA model.

The Generalized Autoregressive Conditionally Heteroscedastic model of order m and r, abbreviated as GARCH(m, r), is of the form

$$w_t = \sigma_t \varepsilon_t \tag{4}$$

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{m} \alpha_j w_{t-j}^2 + \sum_{j=1}^{r} \beta_j \sigma_{t-j}^2, \tag{5}$$

where w_t is noise, i.e. the residual of the ARMA model with mean zero, σ_t^2 is the variance of the noise, ε_t is standard Gaussian white noise $\varepsilon_t \sim$ iidN(0, 1); α_0, α_1, α_2, ..., α_m, β_1, β_2, ..., β_r are constants ($\alpha_0 > 0$, $\alpha_j \geq 0$, $\beta_j \geq 0$). The GARCH(m, r) model assumes that the variance σ_t^2 is defined as the sum of the constant α_0, the linear combination of m past values of a squared noise w_t^2 and the linear combination of r past values of variance σ_w^2 (Shumway et al., 2011; Cowpertwait et al., 2009; Bollerslev, 1986).

The flowchart in Fig. 1 illustrates the decomposition of the autocorrelated time series vector x_t in the combined ARMA+GARCH time series model. As can be seen in the Fig. 1, data decomposition is performed consecutively, first in the fitted ARMA model and then in the fitted GARCH model. In the model fitting, model coefficients are determined so that the sum of the squared model estimate errors is minimized. This fitting approach is called the least squares method, abbreviated as LS method. The LS method is commonly used in model fitting and will also be used in this research.

The estimate error, i.e. residual, is the deviation between the data point and the corresponding value estimated using the model. In Fig. 1, estimate errors, i.e. residuals of the fitted ARMA model and the fitted GARCH model are represented by the residual vector w_t and the standard Gaussian white noise vector ε_t.

The plots in Fig. 2, sections (a-d), are based on the simulated time series vector x_t with $2 \cdot 10^3$ observations. They illustrate the behavior of the time series vector x_t, residual vector w_t, the variance vector σ_t^2 and the standard Gaussian white noise vector ε_t.

When comparing the plots in sections (a-b, d), it can be clearly seen that the information content of these vectors decreases along the decomposition steps.

In the first decomposition, the ARMA model is fitted to the time series vector x_t. As illustrated in Fig. 1, the fitted model explains the AR effect by decomposing the data to coefficients ϕ_1, ϕ_2, ..., ϕ_p (1).

The MA effect is explained by decomposing the data to coefficients θ_1, θ_2, ..., θ_q (2). Unlike the time series vector x_t, the residual vector w_t is not autocorrelated because the autocorrelation effect has already been decomposed to the AR and MA coefficients. However, if a squared residual vector i.e. w_t^2 is autocorrelated, the residual vector w_t contains information on varying variance of the time series vector x_t, which could not be explained using the fitted ARMA model.

The second decomposition is performed for the residual vector w_t in order to explain its varying variance information. Thus the GARCH model is fitted to the residual vector w_t. The fitted GARCH model decomposes information on varying variance to the coefficients α_0, α_1, α_2, ..., α_m, β_1, β_2, ..., β_r. Consequently, the standard Gaussian white noise vector ε_t contains no information as can be seen in Fig. 2, section (d), because $\varepsilon_t \sim$ iidN(0, 1). Therefore the values of the vector ε_t are independent and identically and normally distributed with mean zero and variance one. In addition, the standard Gaussian white noise vector ε_t and its square are not autocorrelated because all the remaining information in the residual vector w_t is decomposed to the coefficients of the fitted GARCH model. The variance vector σ_t^2 in Fig. 2, section (c), is determined based on these coefficients (5).

As mentioned before, the ARMA model cannot explain the varying variance included in the time series vector x_t (Shumway et al., 2011; Box et al., 1976). Consequently, the residual vector w_t contains unexplained varying variance information so it makes variance behavior of the time series vector x_t clearly visible. Therefore, the variance vector σ_t^2 in Fig. 2, section (c), illustrates the variance behavior of the time series vector x_t although it is determined on the basis of the residual vector w_t.

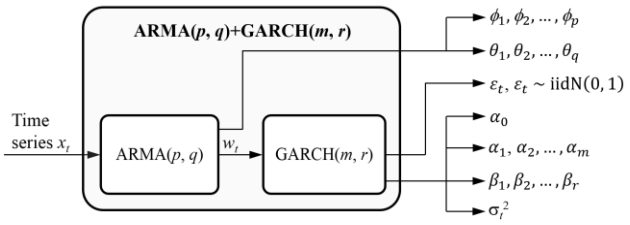


Figure 1. Decomposition of time series vector x_t to residual vector w_t, variance vector σ_t^2, standard Gaussian white noise vector ε_t and coefficients ϕ_1, ϕ_2, …, ϕ_p, θ_1, θ_2, …, θ_q, α_0, α_1, α_2, …, α_m, β_1, β_2, …, β_r.

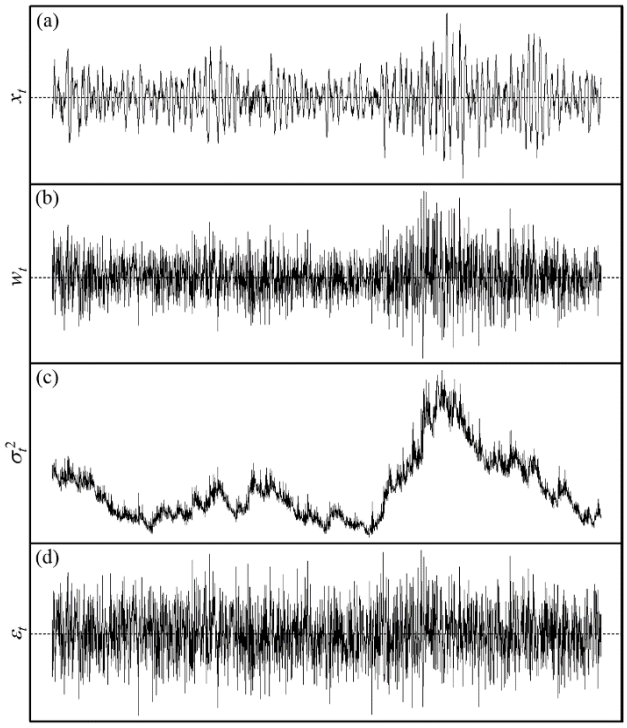


Figure 2. (a) Time series vector x_t, (b) residual vector w_t, (c) variance vector σ_t^2 and (d) standard Gaussian white noise vector ε_t.

2.3 Detection Limit and Detection Sensitivity

An incipient process failure causing abnormally high values in the variance vector σ_t^2 is challenging to identify, because the variance also fluctuates inherently in the stable process condition. Based on simulations, the empirical distributions of the variance vector values were not normal. Therefore, the Statistical Process Control charts, abbreviated SPC charts, for identifying abnormally high variance values could not be used due to their underlying assumptions of normality (Montgomery, 2009).

In case the empirical distribution of variance vector values is not normal, abnormally high variance values can be detected when they exceed a quantile of the empirical distribution (Montgomery, 2009). The empirical distribution of the variance vector σ_t^2 is defined based on the process period that is known to be stable. The quantile is set by an adjustable detection limit, abbreviated as p_k. The quantile, later called p_k quantile, divides an empirical distribution so that p_k proportion of the variance values is smaller than the p_k quantile. $p_k \in [0, 1]$. When the objective is to detect abnormally high variance values, the detection limit p_k is set close to one.

Although the variance vector σ_t^2 values are at a reasonably low level, they may still have individual high peak values, i.e. outliers triggering a false alarm of an abnormally high variance. The variance vector may also have an ascending trend while an individual high peak triggers a premature alarm even though the general variance level is below the p_k quantile.

The number of false alarms can be reduced using adjustable detection sensitivity, abbreviated as *s*, which specifies the minimum number of the consecutive values of the variance vector σ_t^2 above the p_k quantile before the alarm is triggered.

3 Case Study and Simulation

This chapter introduces a case study conducted in the papermaking industry. Objective was to detect a problematic air leakage as early as possible in a paper machine. Steps of the combined ARMA+GARCH time series model use for an incipient air leakage detection are examined. Ability of the examined method to detect air leakage with minimum delay was tested by simulation.

3.1 Background and the Process Data

An industrial case study was conducted for a paper machine, where it is critical to reliably detect an incipient air leakage in a certain vacuum chamber. Based on experience, air leakage in the vacuum chamber gradually increases variance in the chamber pressure while its mean still remains constant. When air leakage increases, the pressure variance also increases until the pressure value exceeds its tolerance limit. Crossing the tolerance limit of the vacuum pressure triggers a sequence of certain mechanical functions in the process, causing an audible sound that is currently subjectively interpreted as a sign of air leakage.

Fig. 3 illustrates the behavior of the pressure in the vacuum chamber. The pressure data used in the case study were recorded at 8 s sampling intervals during a 140 h period. The pressure is stationary, i.e. it varies around the constant mean. Therefore, there is no need for differencing to make it compatible with the ARMA model (Shumway et al., 2011). Pressure variance slightly increases at 70 h of the time. Although the variance increase is detectable, it may be caused by an inherent fluctuation of the pressure variance in a stable and faultless process condition.

3.2 Fitting of a Combined ARMA+GARCH Model

In order to define a typical behavior of the chamber pressure in terms of the coefficients of the combined ARMA+GARCH model, the model is fitted to such a period of the pressure data that represents the stable and faultless process. According to an interview of the paper machine specialists, there were no air leakages observed during the data collection. Thus the pressure behavior including the variance increase at 70 h of the time, as shown in Fig. 3, is considered to be typical for the process. Therefore, the combined ARMA+GARCH model was fitted to the data of the whole 140 h period.

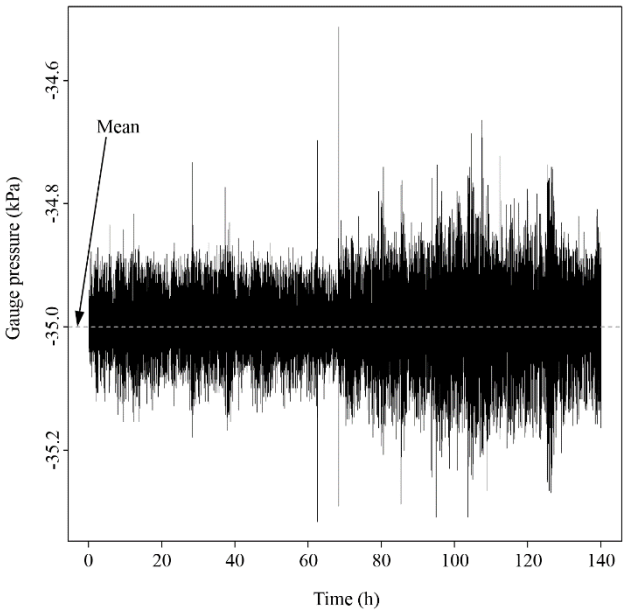


Figure 3. Gauge pressure of the vacuum chamber in the paper machine. Data were recorded at 8 s sampling intervals during a 140 h period.

The best fitting to the data was found when the combined ARMA(4, 4)+GARCH(1, 5) model was used. This means that the ARMA model orders are $p = 4$, $q = 4$ and the GARCH model orders are $m = 1$, $r = 5$.

Based on the definitions of the ARMA model (3) and GARCH model (4, 5), the combined ARMA(4, 4) +GARCH(1, 5) model is of the form

$$\begin{aligned} x_t = &-0.6790 + 2.1828 \cdot x_{t-1} - 1.1515 \cdot x_{t-2} \\ &- 0.3733 \cdot x_{t-3} + 0.3226 \cdot x_{t-4} + \sigma_t \varepsilon_t \\ &- 1.6268 \cdot \sigma_{t-1}\varepsilon_{t-1} + 0.3339 \cdot \sigma_{t-2}\varepsilon_{t-2} \\ &+ 0.4576 \cdot \sigma_{t-3}\varepsilon_{t-3} - 0.1537 \cdot \sigma_{t-4}\varepsilon_{t-4} \end{aligned} \quad (6)$$

$$\begin{aligned} \sigma_t^2 = &\ 0.000001 + 0.036782 \cdot (\sigma_{t-1}\varepsilon_{t-1})^2 \\ &+ 0.323369 \cdot \sigma^2_{t-1} + 0.097586 \cdot \sigma^2_{t-4} \\ &+ 0.541243 \cdot \sigma^2_{t-5} \end{aligned} \quad (7)$$

The terms of the equation (7), which are related to σ^2_{t-2} and σ^2_{t-3}, were omitted because their coefficients were not statistically significant, i.e. zero was included in their 95% confidence interval.

3.3 Air Leakage Simulation

Since no air leakage occurred in the vacuum chamber during the data collection, it had to be created artificially through simulation. Fig. 4, sections (a-c), illustrate the simulation steps from left to right in the same y-axis scale. At the beginning, in Fig. 4, section (a), the time series vector x_t with $5{\cdot}10^3$ observations is simulated according to the combined ARMA(4, 4)+GARCH(1, 5) model that was determined in the previous chapter. The simulated time series vector x_t represents the typical behavior of vacuum chamber pressure in the stable and faultless process condition.

A disturbance vector with $5{\cdot}10^3$ observations is simulated to mimic an incipient and gradually increasing air leakage in the vacuum chamber. The first $2{\cdot}10^3$ observations are constant zero, due to the initial phase of the process that is not disturbed. Fig. 4, section (b), illustrates with a vertical dashed line how since the start of the leakage the variance of the normally distributed random vector linearly increases from zero to its maximum, which is set to low in order to make the disturbance challenging to detect.

The simulated vector in Fig. 4, section (a), and the disturbance vector in section (b) are summed. The sum vector, illustrated in section (c) mimics a pressure of the vacuum chamber in an incipient and gradually increasing air leakage condition that starts after $2{\cdot}10^3$ observations.

The sum vector in Fig. 4, section (c), represents the time series vector x_t as in Fig. 2, section (a), and is decomposed in the fitted combined ARMA(4, 4) +GARCH(1, 5) model. In the decomposition, the variance vector σ_t^2 is determined (5) based on the coefficients of the fitted GARCH(1, 5) model. The steps described in this chapter are repeated in each simulation round.

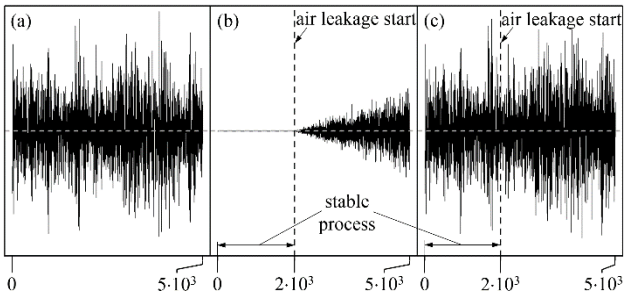


Figure 4. (a) Simulated pressure vector representing the stable and faultless process that follows the combined ARMA(4, 4)+GARCH(1, 5) model. (b) Disturbance vector representing an increasing air leakage. (c) Sum vector that includes the simulated pressure vector and the disturbance vector.

3.4 Air Leakage Detection

Air leakage, i.e. an abnormally high variance in vacuum chamber pressure, should be detected from the variance vector σ_t^2 with as short a delay as possible after the start of the air leakage. Fig. 5 illustrates the behavior of the variance vector σ_t^2 in one simulation round. A vertical dashed line illustrates a time point when an incipient and gradually increasing air leakage starts, as in Fig. 4, section (b). Before the start of the air leakage, the variance vector σ_t^2 represents the stable process. That part of the vector is used for defining the detection threshold of an abnormally high variance. In the threshold setting, the p_k quantile was defined by setting the detection limit $p_k = 0.999$.

Thus only 1‰ of the variance vector observations during the stable process are above the p_k quantile. In case it occurs, it is considered a rare event and therefore an abnormally high variance observation.

In Fig. 5, after some delay since the start of the air leakage, variance vector σ_t^2 adopts an ascending trend until its first values cross the pk quantile. As can be seen, the first crossings of the p_k quantile do not trigger an alarm of an abnormally high variance due to detection sensitivity *s*. As soon as *s* pcs of the consecutive observations above the p_k quantile are counted, an alarm for an abnormally high variance is triggered and a detection delay, abbreviated as d_i, of the simulation round *i* is recorded. d_i describes the delay counted from the start of the air leakage until its detection as the number of variance vector σ_t^2 observations.

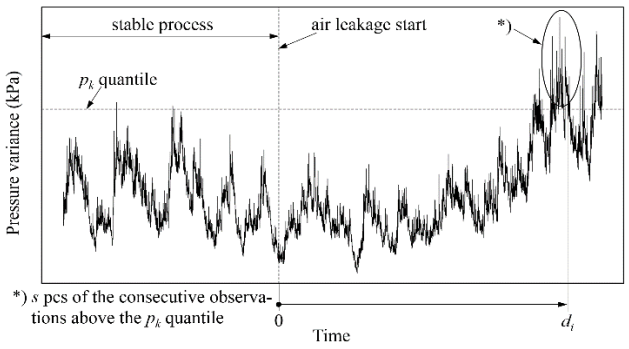


Figure 5. Vector of pressure variance in an air leakage simulation round *i*, $i = 1, 2, \ldots, N_{sim}$. Air leakage starts at time point zero at vertical dashed line. Pressure variance σ_t^2 increases towards the right until the air leakage is detected in delay d_i, when *s* pcs of the consecutive variance vector observations above the p_k quantile are counted.

The detection delay d_i is recorded in each simulation round *i*, $i = 1, 2, \ldots, N_{sim}$, In this study, the $N_{sim} = 27 \cdot 10^3$. The detection delays of all the simulation rounds were summarized in an empirical cumulative probability curve illustrated in Fig 6, section (a). The cumulative probability curve refers the probability to detect an air leakage using the given delay.

Detection sensitivity *s* has an effect onto the shape and location of the cumulative probability curve. The lower the *s* is, the higher is the probability to detect an air leakage using the given delay, i.e. the detection becomes more sensitive. On the other hand, the lower the *s* is, the greater is the type I error (Montgomery, 2009), i.e. the greater is the likelihood of a false air leakage alarm in the stable process phase when there is actually no air leakage. Based on the simulation data, the detection sensitivity was set to $s = 60$ in order to limit the maximum of the type I error rate to 10%. If the type I error rate needs to be lowered, it reduces the sensitivity of the air leakage detection. Thus the compromise between these two depends on the cost of making type I error and the cost of not being able to detect air leakage on time.

Fig. 6, section (a), illustrates that when controlling the variance vector σ_t^2 of the combined ARMA(4, 4) +GARCH(1, 5) model, half of the air leakages were detected with a delay of 2214 observations, whereas four out of five air leakages were detected with a delay of 2875 observations.

In order to benchmark the studied detection method with regard to the alternative one, the detection of the same simulated air leakages was tested using a moving variance vector that was defined directly from the pressure vector in each simulation round *i*. The window size of the moving variance was 25, and it moved stepwise along the pressure vector with step size of 25. Thus, unlike the variance vector σ_t^2 of the ARMA +GARCH model, the resulting variance vector had 200 observations where the stable process phase is represented by the 80 observations that were used for the p_k quantile definition. To make the detection delays d_i comparable, they were multiplied by 25 to transform them back to the same x-axis scale with the Fig. 6, section (a). The cumulative probability curve as illustrated in Fig. 6, section (b), curve (ii), was drawn based on the transformed detection delays.

Fig. 6, section (b), illustrates the cumulative probabilities of the studied detection method with curve (i) and the benchmarked detection method with curve (ii). Their comparison shows that a maximum 56% of the simulated air leakages were detected faster, i.e. with a shorter delay, when the benchmarked method was used. Methods are equally effective for detecting an air leakage with a delay of 2324 observations. When detection delay exceeds 2324 observations, the studied method becomes more effective.

For the delay of 3000 observations, the studied method failed to detect 16% of the simulated air leakages whereas the benchmarked method could not detect 37% of the simulated air leakages. This can be seen from the endpoint of curves (i) and (ii) in the Fig. 5, section (b).

Fig. 6, section (b), shows that the studied method is slower but more reliable than the benchmarked one, though reducing the type I error rate of the studied method impairs its detection speed and reliability.

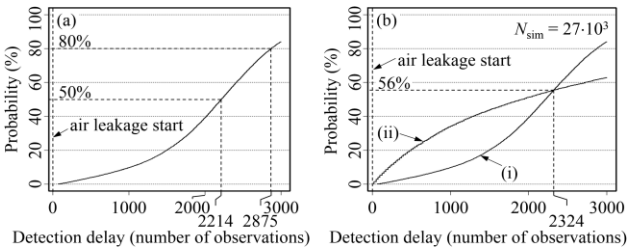


Figure 6. (a) Cumulative probability curve of the detection delay d_i when detection is based on the variance vector σ_t^2 of the combined ARMA(4, 4)+GARCH(1, 5) model. (b) Comparison of the cumulative probability curves: Curve (i) same as in section (a), curve (ii) when detection is based on the moving variance vector that is defined as a moving variance of the simulated pressure vector.

4 Conclusions

An abnormally high variance increase in the critical process variable is often a sign of an incipient failure or other serious disturbance in the process. In this research, the use of the variance vector created using the combined ARMA+GARCH time series model was investigated to detect an abnormal variance increase reliably and with a minimum detection delay. The combined ARMA+GARCH model decomposes a critical variable signal to the standard Gaussian white noise signal and coefficients that determine the AR, MA and the GARCH effects. The variance vector is created based on the coefficients of the GARCH effect. It provides clear visibility to the variance behavior of the critical process variable due to the removed AR and MA effects that disturb the visibility of the variance behavior. The detection of abnormally high values in the variance vector was carried out by setting a detection limit based on the process phase that is known to be stable and faultless.

A case study was conducted about the critical process variable, i.e. the pressure of the vacuum chamber in the paper machine. Process disturbance, i.e. air leakage in the vacuum chamber, was simulated due to its absence during the pressure measurement. The measured pressure data were used as a basis of the simulation. The simulation results proved that the studied method is somewhat slower but more reliable than the benchmarked one where the controlled variance vector is based on the stepwise moving variance of the critical process variable. The studied method failed to detect 16% of the simulated air leakages whereas the benchmarked method failed to detect 37% of the simulated air leakages.

The combined use of the ARMA and GARCH time series models for the Statistical Process Control in industry applications is a fruitful area for further research. When applying these methods in practice, it is beneficial to decompose the data close to their sources using combined ARMA+GARCH time series modelling and to broadcast the decomposition results onward.

References

T. Bollerslev. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986. doi:10.1016/0304-4076(86)90063-1

G.E.P. Box and G.M. Jenkins. *Time Series Analysis, Forecasting, and Control. Revised Edition*, pages 46–84. Oakland, CA: Holden-Day, 1976.

W. Caesarendra, A. Widodo, P.H. Thom, B.-S. Yang and J.D. Setiawan. Combined Probability Approach and Indirect Data-Driven Method for Bearing Degradation Prognostics. *IEEE Transactions on Reliability*, 60(1):14–20, March, 2011. doi:10.1109/TR.2011.2104716

C. Chen, L.Yu and L. Chen. Structural nonlinear damage detection based on ARMA-GARCH model. *Applied Mechanics and Materials*, volumes 204–208, pages. 2891–2896, 2012. doi:10.4028/www.scientific.net/AMM.204-208.2891

P.S.P. Cowpertwait and A.V. Metcalfe. *Introductory Time Series with R.* Springer Science+Business Media, page 148. New York, 2009.

S. Du, L. Xi, J. Yu and J. Sun. Online intelligent monitoring and diagnosis of aircraft horizontal stabilizer assembly processes. *International Journal of Advanced Manufacturing Technology*, 50(1):377–389, 2010. doi:10.1007/s00170-009-2490-0

J. Fan and Q. Yao. *Nonlinear Time Series, Nonparametric and Parametric Methods*, Springer Science+Business Media, pages 10–13, New York, 2003.

V.K. Jandhyala, S.B. Fotopoulos and D.M. Hawkins. Detection and estimation of abrupt changes in the variability of a process. *Computational Statistics & Data Analysis*, 40(1):1–19, 2002. doi:10.1016/S0167-9473(01)00108-6

F. Li, L. Ye, G. Zhang and G. Meng. Bearing Fault Detection Using Higher-Order Statistics Based ARMA Model. *Key Engineering Materials*, volume 347, pages 271–276, 2007. doi:10.4028/0-87849-444-8.271

C.M. Meyer and J.F. Zakrajsek. Rocket Engine Failure Detection Using System Identification Techniques. *AIAA, SAE, ASME, and ASEE, Joint Propulsion Conference, 26th*, Orlando, FL, UNITED STATES, July, 1990.

D.C. Montgomery. *Statistical Quality Control A Modern Introduction, Sixth Edition, International Student Version*. John Wiley & Sons, pages 113 and 264, New York, 2009.

H.T. Pham and B.-S. Yang. Estimation and forecasting of machine health condition using ARMA/GARCH model. *Mechanical Systems and Signal Processing*, 24(2):546–558, 2010. doi:10.1016/j.ymssp.2009.08.004

R. Rzadkowski, E. Rokicki, L. Piechowski and R. Szczepanik. Analysis of middle bearing failure in rotor jet engine using tip-timing and tip-clearance techniques. *Mechanical Systems and Signal Processing*, volumes 76–77, pages 213–227, 2016. doi:10.1016/j.ymssp.2016.01.014

R.H. Shumway and D.S. Stoffer. *Time Series Analysis and Its Applications: With R Examples, Third Edition*, Springer Science+Business Media, pages 83–93,141,286, New York, 2011.

X. Tao, J. Xu, L. Yang and Y. Liu. Bearing fault diagnosis with MSVM based on a GARCH model. *Zhendong yu Chongji/Journal of Vibration and Shock*, 29(5):11–15, May, 2010.

G. Wang, Z. Luo, X. Qin, Y. Leng and T. Wang. Fault identification and classification of rolling element bearing based on time-varying autoregressive spectrum. *Mechanical Systems and Signal Processing*, 22(4):934–947, 2008. doi:10.1016/j.ymssp.2007.10.008

W. Wang and A.K. Wong. Autoregressive Model-Based Gear Fault Diagnosis. *Journal of Vibration and Acoustics*, 124(2):172–179, April, 2002. doi:10.1115/1.1456905

L. Zhao, W. Yu and R. Yan. Rolling Bearing Fault Diagnosis Based on CEEMD and Time Series Modeling. *Mathematical Problems in Engineering*, volume 2014, 2014. Article ID 101867. doi:10.1155/2014/101867

Modeling and Simulation of Train Networks using MaxPlus Algebra

Hazem Al-Bermanei[1] Jari M. Böling[2] Göran Högnäs [3]

[1]Faculty of Business ICT and Life Sciences, Turku University of Applied Sciences, Turku, Finland, hazem.al-bermanei@turkuamk.fi.
[2]Department of Chemical Engineering, Åbo Akademi University, Turku, Finland, jboling@abo.fi.
[3]Department of Mathematics and Statistics, Åbo Akademi University, Turku, Finland

Abstract

Max-plus algebra provides mathematical methods for solving nonlinear problems that can be given the form of linear problems. Problems of this type, sometimes of an administrative nature, arise in areas such as manufacturing, transportation, allocation of resources, and information processing technology. Train networks can be modelled as a directed graph, in which nodes correspond to arrivals and departures at stations, and arcs to travelling times. A particular difficulty is represented by meeting conditions in a single-track railway system. Compared to earlier work which typically include numerical optimization, max-plus formalism is used throughout this paper. The stability and sensitivity of the timetable is analyzed, and different types of delays and delay behavior are discussed and simulated. Interpretation of the recovery matrix is also done. A simple train network with real world background is used for illustration.

Keywords: train schedules, meeting conditions, max-plus algebra, discrete-event systems, delay sensitivity, recovery matrix

1. Introduction

The increasingly saturated European railway infrastructure has, among other concerns, drawn attention to the stability of train schedules as they may cause of domino effect delays across the entire network. A train timetable must be insensitive with regard to small disturbances so that recovery from such disturbances can occur without external control. After a break of self-regulation, this behavior schedule requires the distribution of accurate recovery times and buffer times to reduce delays and prevent the propagation of delay, respectively. Schedule models for railways are usually based on deterministic process times (running times, and transfer times). Moreover, running times are rounded and train tracks are modified to fit the schedule or constraints. The validity of these decisions and streamline schedules must be evaluated to ensure the viability and stability and durability, with respect to network mutual relations and differences in process times. Train networks can be modeled using max-plus algebra (D'Ariano et al., 2007). Stability can be evaluated by calculating the eigenvalue of the matrix in max-plus algebra (Baccelli et al., 1992; van den Boom and De Schutter, 2004; van den Boom et al., 2012; Corman et al., 2012). This eigenvalue is the minimum cycle time required to satisfy all of the schedule and progress constraints, where the timetable operating with this eigenvalue time is given by the associated eigenvector (Baccelli et al., 1992; De Schutter and van den Boom, 2008). Thus, if the eigenvalue λ is more than the intended length of the schedule T, then the schedule is unstable. If $\lambda<T$ the schedule will be stable, and critical if $\lambda=T$ (van den Boom et al., 2012; Corman et al., 2012).

If individual trains are delayed, the effect on the whole network is quite difficult to predict. Smaller delays can typically be absorbed by speeding up the trains, and this can be handled by using max-plus algebra. Larger delays are often handled by rescheduling, typically using optimization, see for example De Schutter et al., (2002); D'Ariano et al., (2007); Corman et al., (2012); and van den Boom and De Schutter, (2004).

In this paper we study the impact of both permanent and dynamic delays in a train network, but restrict ourselves to using max-plus algebra, and thus we do not consider rescheduling. So in practice our study is limited to delays up to half of the cycle time. Meeting conditions, including those introduced by having single tracks, are also fully handled using max-plus state-space formalism, by extending the state with delayed states. When constructing a recovery matrix (van den Boom et al., 2012). Of this extended system, it naturally results in redundancy, as the same physical state appears many times. This redundant recovery information can however be incorrect, due to that no constraints are specified for the delayed states, which are only shifted copies of the most recent state. The parts of the recovery matrix corresponding to the most recent states are still valid.

2. Max-plus algebra

In max-plus algebra we work with the max-plus semi-ring which is the $\mathbb{R}_{max} = \mathbb{R} \cup \{-\infty\}$ and the two binary operations addition $\oplus$ and multiplication $\otimes$, which are defined by:
$a \oplus b = \max(a, b)$, $a \otimes b = a + b$, and $(-\infty) + a = -\infty$.
Define $\varepsilon = -\infty$ and $e = 0$. The additive and multiplicative identities are thus ε and e respectively and the operations are associative, commutative and distributive as in conventional algebra. Furthermore the pair of operations $(\oplus, \otimes)$ can be extended to matrices and vectors similarly as in conventional linear algebra:

- For all $A, B \in \mathbb{R}_{max}^{m \times n}$, $(A \oplus B)_{ij} = a_{ij} \oplus b_{ij} = \max(a_{ij}, b_{ij})$
- For $A \in \mathbb{R}_{max}^{m \times n}$ and $B \in \mathbb{R}_{max}^{m \times n}$ define their product by
$$(A \otimes B)_{ij} = \bigoplus_{j=1}^{k} (a_{ij} \otimes b_{ij}) = \max_{j=\{1,2,\ldots,k\}} (a_{ij} + b_{ij}) \quad 1 \le i \le m, 1 \le l \le p$$
The $n \times n$ identity matrix I_n in max-plus is defined as:
$$I_n = \begin{cases} e & \text{if } i = j \\ \varepsilon & \text{if } i \neq j \end{cases}$$
For $A \in \mathbb{R}_{max}^{m \times n}$, $I_m \otimes A = A \otimes I_n = A$
- For a square matrix A and positive integer n the n^{th} power of A is written as: $A^{\otimes n}$ and it is defined by
$$A^{\otimes n} = \underbrace{A \otimes A \otimes \ldots\ldots\ldots\ldots\ldots \otimes A}_{n \text{ times}}$$

See also Heidegrott et al. (2006), De Schutter and van den Boom (2004), and Baccelli et al. (1992).

3. Scheduled max-plus linear systems

Consider the train network in Figure 1 (vr.fi, 2014). This is a simple network consisting of four stations, Helsinki (H), Karjaa (K), Salo (S) and Turku (T). The end stations are modeled with nodes for both arrival (A in front of the city first letter) and departure (D). The stops at the intermediate stations are short, and thus only the departures are modeled. The weights d_i on the arcs corresponds to the traveling times, while d_1 and d_5 are service times at end stations. The stations between Helsinki and Karjaa are connected by double tracks, and the other connections are single tracks that introduce meeting time conditions. There are five trains available for this, which also introduce some constraints:

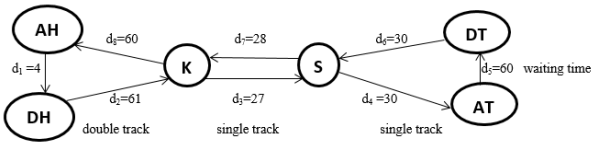


Figure 1. The railroad network between Helsinki and Turku in Finland.

Table 1 provides the schedule (vr.fi, 2014) of five trains running regularly between Helsinki and Turku, and gives the information in connection with the nominal travelling times and the departures.

Table 1. Train time table for trains 1, …, 5 between Turku and Helsinki in hours: minutes. Abbreviations: D=departure, A=arrival, T=Turku, and H=Helsinki.

	1	2	3	4	5
DH	8:02	9:02	10:02	12:02	11:02
AT	10:00	11:00	12:00	14:00	13:00
DT	11:00	12:00	13:00	15:00	14:00
AH	12:58	13:58	14:58	16:58	15:58
DH	13:02	14:02	15:02	17:02	16:02
AT	15:00	16:00	17:00	19:00	18:00
DT	16:00	17:00	18:00	20:00	19:00
AH	17:58	18:58	19:58	21:58	20:58

Now, in order to define the train network as a discrete event system (DES), a state vector is defined as $\boldsymbol{x} = (\boldsymbol{x_{DH}}, \boldsymbol{x_{KS}}, \boldsymbol{x_{ST}}, \boldsymbol{x_{AT}}, \boldsymbol{x_{DT}}, \boldsymbol{x_{SK}}, \boldsymbol{x_{KH}}, \boldsymbol{x_{AH}})^{\boldsymbol{T}}$

with descriptive subscripts:

DH=	departure from Helsinki	DT=	departure from Turku
KS=	departure from Karjaa to Salo	SK=	Salo to Karjaa
ST=	Salo to Turku	KH=	Karjaa to Helsinki
AT=	arrival to Turku	AH=	arrival to Helsinki

The argument *k* on the states denote the kth departure or when indicated arrival to the end stations. Furthermore *k* also indicates the train number, so that $x_{DH}(k)$ is the departure time from Helsinki for train *k*, and $x_{AT}(k)$ is the arrival of the same train to Turku.

The period of the time table is T=60 minutes. Due to that we have only one track between Karjaa and Salo, and between Salo and Turku, we get the following meeting conditions:

$x_{KS}(k) \geq x_{KH}(k-3)$ (in Karjaa),

$x_{ST}(k) \geq x_{SK}(k-2)$ (in Salo for the train going towards Turku),

$x_{DT}(k) \geq x_{AT}(k+1)$ (in Turku),

$x_{SK}(k) \geq x_{ST}(k+2)$ (in Salo for the train going towards Karjaa).

Combination of the meeting conditions and the constraints introduced by travelling times gives the following equations (the first one comes from having only five trains):

$$x_{DH}(k) = x_{AH}(k-5) + d_1,$$
$$x_{KS}(k) = \max\big(x_{DH}(k) + d_2, x_{KH}(k-3)\big)$$
$$x_{ST}(k) = \max\big(x_{KS}(k) + d_3, x_{SK}(k-2)\big)$$
$$x_{AT}(k) = x_{ST}(k) + d_4$$
$$x_{DT}(k) = \max\big(x_{AT}(k) + d_5, x_{AT}(k+1)\big) \quad (1)$$
$$x_{SK}(k) = \max\big(x_{DT}(k) + d_6, x_{ST}(k+2)\big) \quad (2)$$
$$x_{KH}(k) = x_{SK}(k) + d_7$$
$$x_{AH}(k) = x_{KH}(k) + d_8$$

In order to get an equation of type $x(k) = A \otimes x(k-1)$, the right hand side expressions containing *k* or higher indices are substituted with expressions containing index *k*-1 at most:

$$x_{DH}(k) = x_{AH}(k-5) + d_1,$$

$$x_{KS}(k) = \max\big(x_{AH}(k-5) + d_1 + d_2, x_{KH}(k-3)\big),$$

$$x_{ST}(k) = \max(x_{AH}(k-5) + d_1 + d_2 + d_3,\ x_{SK}(k-2),\ x_{KH}(k-3) + d_3),$$

$$x_{AT}(k) = \max(x_{AH}(k-5) + d_1 + d_2 + d_3 + d_4,\ x_{SK}(k-2) + d_4,\ x_{KH}(k-3) + d_3 + d_4),$$

$$x_{DT}(k) = \max(x_{AH}(k-5) + d_1 + d_2 + d_3 + d_4 + d_5, x_{SK}(k-2) + d_4 + d_5, x_{KH}(k-3) + d_3 + d_4 + d_5, x_{AH}(k-4) + d_1 + d_2 + d_3 + d_4,\ x_{SK}(k-1) + d_4,\ x_{KH}(k-2) + d_3 + d_4),$$

$$x_{SK}(k) = \max(x_{AH}(k-5) + d_1 + d_2 + d_3 + d_4 + d_5 + d_6, x_{SK}(k-2) + d_4 + d_5 + d_6,\ x_{KH}(k-3) + d_3 + d_4 + d_5 + d_6, x_{AH}(k-4) + d_1 + d_2 + d_3 + d_4 + d_6,\ x_{SK}(k-1) + d_4 + d_6,\ x_{KH}(k-2) + d_3 + d_4 + d_6, x_{AH}(k-3) + d_1 + d_2 + d_3,\ x_{KH}(k-1) + d_3),$$

$$x_{KH}(k) = \max(x_{AH}(k-5) + d_1 + d_2 + d_3 + d_4 + d_5 + d_6 + d_7, x_{SK}(k-2) + d_4 + d_5 + d_6 + d_7, x_{KH}(k-3) + d_3 + d_4 + d_5 + d_6 + d_7, x_{AH}(k-4) + d_1 + d_2 + d_3 + d_4 + d_6 + d_7,\ x_{SK}(k-1) + d_4 + d_6 + d_7,\ x_{KH}(k-2) + d_3 + d_4 + d_6 + d_7, x_{AH}(k-3) + d_1 + d_2 + d_3 + d_7,\ x_{KH}(k-1) + d_3 + d_7),$$

$$x_{AH}(k) = \max(x_{AH}(k-5) + d_1 + d_2 + d_3 + d_4 + d_5 + d_6 + d_7 + d_8, x_{SK}(k-2) + d_4 + d_5 + d_6 + d_7 + d_8, x_{KH}(k-3) + d_3 + d_4 + d_5 + d_6 + d_7 + d_8, x_{AH}(k-4) + d_1 + d_2 + d_3 + d_4 + d_6 + d_7 + d_8,\ x_{SK}(k-1) + d_4 + d_6 + d_7 + d_8,\ x_{KH}(k-2) + d_3 + d_4 + d_6 + d_7 + d_8,\ x_{AH}(k-3) + d_1 + d_2 + d_3 + d_7 + d_8,\ x_{KH}(k-1) + d_3 + d_7 + d_8).$$

Define the augmented system $x_j(k)$ where $j = 1,2,3,\ldots,40$:

$$\left.\begin{aligned}
x_j(k) &= x_{DH}(k-j+1), && j = 1,\ldots,5\\
x_j(k) &= X_{KS}(k-j+6), && j = 6,\ldots,10\\
x_j(k) &= x_{ST}(k-j+11), && j = 11,\ldots,15\\
x_j(k) &= x_{AT}(k-j+16), && j = 16,\ldots,20\\
x_j(k) &= x_{DT}(k-j+21), && j = 21,\ldots,25\\
x_j(k) &= x_{SK}(k-j+26), && j = 26,\ldots,30\\
x_j(k) &= x_{KH}(k-j+31), && j = 31,\ldots,35\\
x_j(k) &= x_{AH}(k-j+36), && j = 36,\ldots,40
\end{aligned}\right\} \quad (3)$$

This means that $x_i(k) = x_{i-1}(k-1)$ for $i = 2,3,\ldots\ldots\ldots,40$ except $i = 1,6,11,16,21,26,31$ and 36. The main equations using numbers as subscripts then become as follows:

$$x_1(k) = x_{40}(k-1) + d_1,$$
$$x_6(k) = \max\big(x_{40}(k-1) + d_1 + d_2, x_{33}(k-1)\big),$$
$$x_{11}(k) = \max(x_{40}(k-1) + d_1 + d_2 + d_3, x_{27}(k-1),\ x_{33}(k-1) + d_3),$$
$$x_{16}(k) = \max(x_{40}(k-1) + d_1 + d_2 + d_3 + d_4,\ x_{27}(k-1) + d_4,\ x_{33}(k-1) + d_3 + d_4),$$
$$x_{21}(k) = \max(x_{40}(k-1) + d_1 + d_2 + d_3 + d_4 + d_5, x_{27}(k-1) + d_4 + d_5, x_{33}(k-1) + d_3 + d_4 + d_5, x_{39}(k-1) + d_1 + d_2 + d_3 + d_4,\ x_{26}(k-1) + d_4,\ x_{32}(k-1) + d_3 + d_4),$$
$$x_{26}(k) = \max(x_{40}(k-1) + d_1 + d_2 + d_3 + d_4 + d_5 + d_6, x_{27}(k-1) + d_4 + d_5 + d_6,\ x_{33}(k-1) + d_3 + d_4 + d_5 + d_6, x_{39}(k-1) + d_1 + d_2 + d_3 + d_4 + d_6,\ x_{26}(k-1) + d_4 + d_6,\ x_{32}(k-1) + d_3 + d_4 + d_6, x_{38}(k-1) + d_1 + d_2 + d_3, x_{31}(k-1) + d_3),$$
$$x_{31}(k) = \max(x_{40}(k-1) + d_1 + d_2 + d_3 + d_4 + d_5 + d_6 + d_7, x_{27}(k-1) + d_4 + d_5 + d_6 + d_7, x_{33}(k-1) + d_3 + d_4 + d_5 + d_6 + d_7, x_{39}(k-1) + d_1 + d_2 + d_3 +$$

$$d_4 + d_6 + d_7,\ x_{26}(k-1) + d_4 + d_6 + d_7,\ x_{32}(k-1) + d_3 + d_4 + d_6 + d_7,\ x_{38}(k-1) + d_1 + d_2 + d_3 + d_7,\ x_{31}(k-1) + d_3 + d_7),$$ and

$$x_{36}(k) = \max(x_{40}(k-1) + d_1 + d_2 + d_3 + d_4 + d_5 + d_6 + d_7 + d_8,\ x_{27}(k-1) + d_4 + d_5 + d_6 + d_7 + d_8,\ x_{33}(k-1) + d_3 + d_4 + d_5 + d_6 + d_7 + d_8,\ x_{39}(k-1) + d_1 + d_2 + d_3 + d_4 + d_6 + d_7 + d_8,\ x_{26}(k-1) + d_4 + d_6 + d_7 + d_8,\ x_{32}(k-1) + d_3 + d_4 + d_6 + d_7 + d_8,\ x_{38}(k-1) + d_1 + d_2 + d_3 + d_7 + d_8,\ x_{31}(k-1) + d_3 + d_7 + d_8).$$

If we rewrite the above evolution equations as a max-plus-linear discrete event systems state space model of the form

$$x(k) = A \otimes x(k-1) \quad (4)$$

we obtain a square matrix A of size 40×40. For example the 36th row in the matrix A is:
[ε . . . . . ε 148 208 ε ε ε 115 175 235 ε ε ε ε 180 240 300],
where the entry 148 has column index 26.
The power method (Baccelli et al., 1992; van den Boom and De Schutter, 2004; De Schutter and van den Boom, 2008) is used for finding the eigenvalue λ of the matrix A. The method means repetitive multiplications $x(k) = A \otimes x(k-1) = A^{\otimes k} \otimes x(0)$, and it stops when there are integers $i > j \geq 0$ and a real number c for which
$x(i) = x(j) \otimes c$. The eigenvalue is then given by $\lambda(A) = \frac{c}{i-j}$. In this case, using $x(0) = \mathbf{0}$, iteration according Equation 2 gives

$x(12) = A \otimes x(11) = [664\ 604\ 544\ 484\ 424\ 725\ 665\ 605\ 545\ 485\ 52\ 692\ 632\ 572\ 512\ 782\ 722\ 662\ 602\ 542\ 842\ 782\ 722\ 662\ 602\ 872\ 812\ 752\ 692\ 632\ 900\ 840\ 780\ 720\ 660\ 960\ 900\ 840\ 780\ 720]^T$,

$x(13) = A \otimes x(12)$
$= [724\ 664\ 604\ 544\ 484\ 785\ 725\ 665\ 605\ 545\ 812\ 752\ 692\ 632\ 572\ 842\ 782\ 722\ 662\ 602\ 902\ 842\ 782\ 722\ 662\ 932\ 872\ 812\ 752\ 692\ 960\ 900\ 840\ 780\ 720\ 1020\ 960\ 900\ 840\ 780\ ]^T$
and
$x(13) = x(12) \otimes 60$

Thus the eigenvalue is $\lambda(A) = 60/(13-12) = 60$. The eigenvalue represents the cycle of the schedule which means that the trains start from each station every 60 minutes.

This also means that $x(13)$ is an eigenvector, and $(x(13) - c)$, where c is any constant, is also an eigenvector. One eigenvector of A is v where

$$v = [0\ \ -60\ \ -120\ \ -180\ \ -240\ \ 61\ \ 1\ \ -59\ \ -119\ \ -179\ \ 88\ \ 28\ \ -32\ \ -92\ \ -152\ \ 118\ \ 58\ \ -2\ \ -62\ \ -122\ \ 178\ \ 118\ \ 58\ \ -2\ \ -6\ \ 208\ \ 148\ \ 88\ \ 28\ \ -32\ \ 236\ \ 176\ \ 116\ \ 56\ \ -2\ \ 296\ \ 236\ \ 176\ \ 116\ \ 56]^T \quad (5)$$

This eigenvector v includes the schedule of the trains, relative to the last departure from Helsinki (the first element of v). So the element -240 means that five departures back a train from Helsinki left 240 minutes ago, and the element 296 means that it takes 296 minutes for a train to come back to Helsinki.

4. Timetable stability

4.1 Delay sensitivity analysis

All the travel times d_i introduced in the Section 3, consist of a minimal travel time and a slack time. Here it is assumed that the minimal travel time is 90% of the nominal time, and the slack is thus 10%. For the small waiting time d_1 in Helsinki it is assumed that there is no slack.

Handling delays is a relevant and common problem in train networks, and the sensitivity of delays can be analyzed using max-plus models. A permanent delay means that the nominal travel times is increased, which is compensated for by decreasing the other travel times to their minimal values. This gives a slightly different system, for which a new eigenvalue can be calculated. The relative and absolute limits for increasing the different travelling times individually without violation of the roundtrip time (i.e. $\lambda > T$) are presented in Table 2.

Table 2: Delay sensitivity of the different traveling times.

Traveling time with delay	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
Relative limit	440 %	18 %	28 %	10 %	10 %	10 %	27.5 %	19.3 %
Absolute limit (min)	17.6	11.5	7.8	3	6	3	7.7	11.6

Table 2 show the maximal value that a single travelling time d_i can be increased, and still get the nominal roundtrip time (given by the eigenvalue of the modified matrix) by decreasing all the other travelling times to their minimal values. For example if we increase d_7 by 27.5% which is equal to 7.7 minutes, and reduce all the other travelling times to their minimal values, we will still get the eigenvalue λ=60.

A limitation with the analysis is that it assumes a permanent change in the delays, and results concerns only steady state. It does not give information about dynamic delay propagation, which is the theme of the following section.

4.2 Dynamic Delay Propagation

The delay sensitivity analysis in Section 4.1 assumed that we had permanent changes in the travelling times. A more normal situation is that the delay only concerns one single travel time, which means that the corresponding max-plus system matrix becomes time varying, due to that the travel times d_i become time varying (indicated by an index k). This is so due to the meeting conditions, that is Equations (1) and (2), where future states $x_{AT}(k+1)$ and $x_{ST}(k+2)$ appear. These are expanded to $\max(x_{AH}(k-4)+d_1(k+1)+d_2(k+1)+d_3(k+1)+d_4(k+1),\ x_{SK}(k-1)+d_4(k+1),\ x_{KH}(k-2)+d_3(k+1)+d_4(k+1))$ and $\max(x_{AH}(k-3)+d_1(k+2)+d_2(k+2)+d_3(k+2),\ x_{KH}(k-1)+d_3(k+2))$ respectively. As indicated with iteration indices newer versions of travel times are needed in these equations. Speeding up can also only be done after the delay has appeared, which in our case means that after a delay in $d_i(k)$ only the traveling times $d_j(k)$ with $j > i$, can be decreased in the same iteration k. In the next iteration all the traveling times can be decreased.

In Table 3 it has been tested how long it takes for a delay of 10, 20 and 30 minutes respectively in a certain travel time, to disappear from the system.

Table 3: Times expressed in minutes that it takes for a delay in a certain traveling time to disappear from the system.

Travel Time	**Delay 10** min	**Delay 20** min	**Delay 30** min
d_1	89.2	182.4	301.3
d_2	88.3	182.4	300.4
d_3	93,2	182.4	300.4
d_4	91	185.1	303.1
d_5	91	185.1	303.1
d_6	93.2	182.4	300.4
d_7	68	184.2	305.3
d_8	89.2	182.4	301.3

The calculation of the disappearance of a delay can be done as follows. Let $M_n$denote a matrix with the nominal timetables, that is $M_n = [v, v\otimes T, v\otimes T^{\otimes 2}, \dots]$, and M_d is a matrix with the delayed arrival and departure times at corresponding times. The part of the time tables that can be used for selecting the part of the time table that is affected by a delay using the logical expression $(M_d - M_n) > 0$. This means that the time instant of the last delay t_d can be found using

$t_d = \max\big[M_d\big((M_d - M_n) > 0\big) - M_n(i,j)\big]$,

where *i, j* are the timetable indices when actual first delay take place. For example 88.3 in on second row second column in Table 3 means that if the single travelling time d_2 is increased by 10 minutes, and the travelling times d_3, d_4, d_5, d_6, d_7 and d_8 are speeded up to their minimal values, then the time instant of the last deviation from the time table is 88.3 minutes after the delay.

4.3 Recovery Matrix

In Goverde (2007) max-plus linear systems are written in polynomial form,

$$x(k) = A_0\otimes x(k)\oplus A\otimes x(k-1)\oplus w(k) \qquad (6)$$

where A is defined as in Equation (4), A_0 is the matrix describing the direct connections from $x(k)$ to $x(k)$, and $w(k)$ is the nominal departure times in period k. A_0 is in this case given by all the direct travelling times d_i, including all delayed states, such that

$A_0(m+5, m) = d_i$, for $m = (i-1)5 + n$, for all $n = 1,2,\dots 5$, and for all $i = 2,3,\dots 8$.

All the other elements of A_0 are ε, as there are no direct connections. The departure times are given by the eigenvector v in Equation 5, and the period T according to $w(k) = T^{\otimes k}\otimes v$. The polynomial equation can be written using a single matrix A_p, according

$$x(k) = A_p\otimes x(k-1)\oplus w(k) \qquad (7)$$

where $A_p = A_0\oplus A\otimes T^{\otimes -1}$.

Definition: Consider the max-plus linear system in Equation (7). The entry r_{ij} of the recovery matrix R is defined as the maximum delay of $x_j(m)$ such that $x_i(k)$ is not delayed for any $k > m$ (Goverde 2007).

The following equation (Baccelli et al., 1992; Goverde 2007) defines the elements of the recovery matrix,

$r_{ij} = w_i - w_j - \left[A_p^+\right]_{ij}$,

where the w_i and w_j are element of vector w, $A_p^+ = \bigoplus_{k=1}^{\infty} A_p^{\otimes k}$, and the notation $\left[A_p^+\right]_{ij}$ refers to the ij^{th} element of the matrix A_p^+. If in the graph of A_p^+ no path exists from node j to node i then $r_{ij} = \infty$. The recovery matrix thus takes values from the extended set $\overline{\mathbb{R}}_{max} = \mathbb{R}_{max} \cup \{\infty\}$.

In the studied train network between Helsinki and Turku, constructed from Table 1 presented in Figure 1, the recovery matrix R is of size 40×40, with $T = 60$. A 20×20 submatrix of that matrix is given in Table 4.

According to Goverde, (2007), the j^{th} column of the recovery matrix *R* gives the recovery time from event *j* to all other events in the timetable and thus represents the impact a delay of event *j* has on future train events, and the i^{th} row of the recovery matrix *R* gives the recovery time from event *i* from all other events in the

timetable and thus represents the sensitivity of event *i* on delays of preceding events. The diagonal elements of *R* again represent recovery times to the event itself. In our example, most of our states are delayed versions of previous states. As can be noted in Table 4, not all diagonal elements representing the same departure at different times are same. For example, $r_{16,16} = 12$, $r_{18,18} = 22.5$ and $r_{19,19} = 29.6$, although these elements all correspond to the event "arrival in Turku" at times k, $k-2$ and $k-3$ respectively. As k is arbitrary, all these recovery elements should logically be the same. This is not so because the delayed versions are just memory variables, for which no other constraints than the back shifting according Equation 3 is present, and thus the recovery matrix is not correct for these. Thus in our example only every fifth row in the recovery matrix show true recovery times, and these are shown in Table 5.

For example the first row in the reduced recovery matrix is easy to interpret; the first value is 29.6, which is the total slack for a single train. After that the slack is reduced by the slack in corresponding travel time, up to the final value 0, which corresponds to that no slack is present in the 4 minute waiting time in Helsinki (d_1). All the other travelling times are assumed to have 10% slack. The other zero (row 11, columns 26) is due to a meeting condition (in Salo).

The results shown in Table 2 can also be calculated using recovery matrix calculations. In Table 2 it was assumed that we have a permanent delay in one travel time. The maximum tolerance for a permanent delay in one travel time can be obtained by increasing the corresponding travel time in the recovery matrix, until we start getting negative entries on the relevant diagonal elements in the recovery matrix (the ones

Table 4: The upper left quadrant of the recovery matrix, with diagonal element shaded.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	29.6	29.6	29.6	29.6	29.6	23.5	23.5	23.5	23.5	23.5	20.8	20.8	20.8	20.8	20.8	17.8	17.8	17.8	17.8	17.8
2	0	29.6	29.6	29.6	29.6	23.5	23.5	23.5	23.5	23.5	20.8	20.8	20.8	20.8	20.8	17.8	17.8	17.8	17.8	17.8
3	0	0	29.6	29.6	29.6	23.5	23.5	23.5	23.5	23.5	20.8	20.8	20.8	20.8	20.8	17.8	17.8	17.8	17.8	17.8
4	0	0	0	29.6	29.6	23.5	23.5	23.5	23.5	23.5	20.8	20.8	20.8	20.8	20.8	17.8	17.8	17.8	17.8	17.8
5	0	0	0	0	29.6	23.5	23.5	23.5	23.5	23.5	20.8	20.8	20.8	20.8	20.8	17.8	17.8	17.8	17.8	17.8
6	6.1	28.6	28.6	35.7	35.7	22.5	22.5	22.5	29.6	29.6	19.8	19.8	19.8	26.9	26.9	16.8	16.8	16.8	23.9	23.9
7	6.1	6.1	28.6	35.7	35.7	0	22.5	22.5	29.6	29.6	19.8	19.8	19.8	26.9	26.9	16.8	16.8	16.8	23.9	23.9
8	6.1	6.1	6.1	35.7	35.7	0	0	22.5	29.6	29.6	19.8	19.8	19.8	26.9	26.9	16.8	16.8	16.8	23.9	23.9
9	6.1	6.1	6.1	6.1	35.7	0	0	0	29.6	29.6	19.8	19.8	19.8	26.9	26.9	16.8	16.8	16.8	23.9	23.9
10	6.1	6.1	6.1	6.1	6.1	0	0	0	0	29.6	19.8	19.8	19.8	26.9	26.9	16.8	16.8	16.8	23.9	23.9
11	8.8	20.8	31.3	38.4	38.4	2.7	14.7	25.2	32.3	32.3	12	12	22.5	29.6	29.6	9	9	19.5	26.6	26.6
12	8.8	8.8	31.3	38.4	38.4	2.7	2.7	25.2	32.3	32.3	0	12	22.5	29.6	29.6	9	9	19.5	26.6	26.6
13	8.8	8.8	8.8	38.4	38.4	2.7	2.7	2.7	32.3	32.3	0	0	22.5	29.6	29.6	9	9	19.5	26.6	26.6
14	8.8	8.8	8.8	8.8	38.4	2.7	2.7	2.7	2.7	32.3	0	0	0	29.6	29.6	9	9	19.5	26.6	26.6
15	8.8	8.8	8.8	8.8	8.8	2.7	2.7	2.7	2.7	2.7	0	0	0	0	29.6	9	9	19.5	26.6	26.6
16	11.8	23.8	34.3	41.4	41.4	5.7	17.7	28.2	35.3	35.3	3	15	25.5	32.6	32.6	12	12	22.5	29.6	29.6
17	11.8	11.8	34.3	41.4	41.4	5.7	5.7	28.2	35.3	35.3	3	3	25.5	32.6	32.6	0	12	22.5	29.6	29.6
18	11.8	11.8	11.8	41.4	41.4	5.7	5.7	5.7	35.3	35.3	3	3	3	32.6	32.6	0	0	22.5	29.6	29.6
19	11.8	11.8	11.8	11.8	41.4	5.7	5.7	5.7	5.7	35.3	3	3	3	3	32.6	0	0	0	29.6	29.6
20	11.8	11.8	11.8	11.8	11.8	5.7	5.7	5.7	5.7	5.7	3	3	3	3	3	0	0	0	0	29.6

Table 5: The relevant parts of the recovery matrix. Diagonal elements highlighted by green, and recovery times related to a full cycle is highlighted with orange

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	29.6	29.6	29.6	29.6	29.6	23.5	23.5	23.5	23.5	23.5	20.8	20.8	20.8	20.8	20.8	17.8	17.8	17.8	17.8	17.8
6	6.1	28.6	28.6	35.7	35.7	22.5	22.5	22.5	29.6	29.6	19.8	19.8	19.8	26.9	26.9	16.8	16.8	16.8	23.9	23.9
11	8.8	20.8	31.3	38.4	38.4	2.7	14.7	25.2	32.3	32.3	12	12	22.5	29.6	29.6	9	9	19.5	26.6	26.6
16	11.8	23.8	34.3	41.4	41.4	5.7	17.7	28.2	35.3	35.3	3	15	25.5	32.6	32.6	12	12	22.5	29.6	29.6
21	17.8	29.8	40.3	41.4	47.4	11.7	23.7	34.2	35.3	41.3	9	21	31.5	32.6	38.6	6	18	28.5	29.6	35.6
26	20.8	32.8	38.4	44.4	50.4	14.7	26.7	32.3	38.3	44.3	12	24	29.6	35.6	41.6	9	21	26.6	32.6	38.6
31	23.6	35.6	41.2	47.2	53.2	17.5	29.5	35.1	41.1	47.1	14.8	26.8	32.4	38.4	44.4	11.8	23.8	29.4	35.4	41.4
36	29.6	41.6	47.2	53.2	59.2	23.5	35.5	41.1	47.1	53.1	20.8	32.8	38.4	44.4	50.4	17.8	29.8	35.4	41.4	47.4
	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	11.8	11.8	11.8	11.8	11.8	8.8	8.8	8.8	8.8	8.8	6	6	6	6	6	0	0	0	0	0
6	10.8	10.8	10.8	17.9	17.9	7.8	7.8	7.8	14.9	14.9	5	5	5	12.1	12.1	6.1	6.1	6.1	6.1	6.1
11	3	3	13.5	20.6	20.6	0	0	10.5	17.6	17.6	7.7	7.7	7.7	14.8	14.8	8.8	8.8	8.8	8.8	8.8
16	6	6	16.5	23.6	23.6	3	3	13.5	20.6	20.6	10.7	10.7	10.7	17.8	17.8	11.8	11.8	11.8	11.8	11.8
21	6	12	22.5	23.6	29.6	3	9	19.5	20.6	26.6	10.7	10.7	16.7	17.8	23.8	11.8	11.8	11.8	11.8	17.8
26	3	15	20.6	26.6	32.6	6	12	17.6	23.6	29.6	7.7	13.7	14.8	20.8	26.8	8.8	8.8	8.8	14.8	20.8
31	5.8	17.8	23.4	29.4	35.4	2.8	14.8	20.4	26.4	32.4	10.5	16.5	17.6	23.6	29.6	11.6	11.6	11.6	17.6	23.6
36	11.8	23.8	29.4	35.4	41.4	8.8	20.8	26.4	32.4	38.4	6	22.5	23.6	29.6	35.6	17.6	17.6	17.6	23.6	29.6

The recovery matrix take in the consideration only one train not the whole system and it gives all the information for the delay of one train only. A 0 in the recovery matrix means a tight schedule, with no slack.

indicated by green in Table 5).

The results in Table 3 can only partially be calculated using recovery matrix calculations. In Table 3 certain temporary delays (10, 20 and 30 minutes) were considered. In Table 3 it can be seen that the time it

takes for the system to catch up after delays of 30 minutes, are all slightly more than 300 minutes. This is not a coincidence, in most cases it is the delayed train itself that uses most time to catch up, and the recovery time 29.6 in positions highlighted with orange, it means that if we have a delay which is larger than 29.6, it will take more than 300 minutes (i.e. a full cycle) for the system to catch up.

5. Conclusions

This paper described how a max-plus model for a train system can be constructed. Meeting conditions caused by having a single track, and other physical constrains, have been handled by extending the state space with delayed states, which has enabled rewriting the state update equation in the form $x(k) = A \otimes x(k-1)$. Static and dynamic delay sensitivity of the network has been analyzed by modifying the A-matrix, and using eigenvalue calculations. The such obtained results were compared to standard recovery matrix based calculations. A recovery matrix for the chosen extended state space becomes large, and contains even irrelevant information. Guidelines for finding and interpreting the relevant information from the recovery matrix have been discussed. Max-plus formalism was used throughout this paper.

References

Andrea D'Ariano, Dario Pacciarelli, Marco Pranzo. A branch and bound algorithm for scheduling trains in a railway network. *European Journal of Operational Research* 183(2):643-657, 2007.

Francois Baccelli, Guy Cohen , Geert Jan Olsder, Jean-Pierre Quadrat. *Synchronization and Linearity - An Algebra for Discrete Event Systems*, Wiley, New York, 1992.

Ton J.J. van den Boom, Bart De Schutter. Modeling and control of railway networks. In proceedings of the *2004 American Control Conference*. Vol. 6, 5728-5733 IEEE, 2004.

Ton J.J. van den Boom, Bart Kersbergen, Bart De Schutter. Structured modeling, analysis, and control of complex railway operations. In Proceedings of the *51st IEEE Conference on Decision and Control*, Maui, Hawaii, 7366-7371, 2012.

Francesco Corman, Andrea D'Ariano, Dario Pacciarelli, Marco Pranzo. Bi-objective conflict detection and resolution in railway traffic management. *Transportation Research Part C: Emerging Technologies* 20(1):79-94, 2012.

Bart De Schutter, Ton J.J. van den Boom. Max-plus algebra and max-plus linear discrete event systems: An introduction. In proceedings of *9th International Workshop on Discrete Event Systems*, IEEE,2008

Bart De Schutter, Ton J.J. van den Boom, Andreas Hegyi. A model predictive control approach for recovery from delays in railway systems. *Transportation Research Record* 1793:15-20, 2002.

Rob M.P. Goverde. Railway timetable stability analysis using max-plus system theory, *Transportation Research Part B: Methodological*, 41(2): 179-201, 2007.

Rob M.P. Goverde. A delay propagation algorithm for large-scale railway traffic networks, *Transportation Research Part C: Emerging Technologies*, 18(3): 269-287, 2010.

Bernd Heidergott, Geert Jan Olsder, Jacob van der Woude. *Max Plus at Work*. Princeton, New Jersey: Princeton University Press, 2006.

Pavle Kecman, Francesco Corman, Andrea D'Ariano, Rob M.P. Goverde. Rescheduling models for railway traffic management in large-scale networks. *Public Transport*, 5(1-2): 95-123, 2013.

Timetables for long-distance trains between Turku and Helsinki, available on https://www.vr.fi/cs/vr/en/long-distance_timetables (accessed 6th of March 2014).

Simulation Metamodeling using Dynamic Bayesian Networks with Multiple Time Scales

Mikko Harju Kai Virtanen Jirka Poropudas

Department of Mathematics and Systems Analysis, Aalto University School of Science, Finland,
mikko.harju@aalto.fi, kai.virtanen@aalto.fi, jirka.poropudas@aalto.fi

Abstract

The utilization of dynamic Bayesian networks (DBNs) in simulation metamodeling enables the investigation of the time evolution of state variables of a simulation model. DBN metamodels have previously described the changes in the probability distribution of the simulation state by using a time slice structure in which the state variables are described at common time instants. In this paper, the novel approach to the determination of the time slice structure is introduced. It enables the selection of time instants of the DBN separately for each state variable. In this way, a more accurate metamodel representing multiple time scales of the variables is achieved. Furthermore, the construction is streamlined by presenting a dynamic programming algorithm for determining the key time instants for individual variables. The construction and use of the DBN metamodels are illustrated by an example problem dealing with the simulated operation of an air base.

Keywords: Bayesian networks, discrete event simulation, dynamic Bayesian networks, simulation, simulation metamodeling

1 Introduction

Discrete event simulation (DES) (e.g. Law and Kelton 2000) is a widely used methodology for modeling and analyzing stochastic dynamic systems. A DES model describes a system consisting of three types of variables (Zeigler et al., 2000). The values of input variables are given prior to the simulation and can be, e.g., parameters that determine the configuration of the system. Time variant state variables describe the time evolution of the system. Output variables obtain values after the simulation is completed and correspond to the characteristics of the system that are being investigated, such as the average waiting time in a queueing system. The main interest in the analysis of DES models is often on the relation between the input and output variables. Simulation metamodels (Friedman, 2012; Kleijnen, 2008) have been used in order to efficiently describe this relationship. See (Poropudas, 2011) for an overview of different types of metamodels as well as details of the construction and utilization of such models.

To better understand how the simulation progresses, it may be of interest to investigate the time development of the state variables of a DES model – instead of the dependence between inputs and outputs. With most simulation metamodeling techniques, such as regression modeling and stochastic Kriging (Kleijnen, 2008), it is not possible to include the state variables into the metamodel. However, the time evolution of the state variables can be analyzed by using dynamic Bayesian networks (DBNs, see, e.g., Murphy 2002) as simulation metamodels (Poropudas and Virtanen, 2007, 2011). Bayesian networks (BNs, see, e.g., Pearl 1986) are probabilistic models that describe the joint probability distribution of discrete random variables. A BN consists of a directed acyclic graph with nodes corresponding to variables and arcs indicating the dependencies between the variables. In addition, a conditional probability table (CPT) is associated with each node, describing its probability distributions conditional on the values of its parent nodes. In a DBN, individual variables are represented by multiple nodes that correspond to their value at specific time instants. In simulation metamodeling, the nodes of a DBN correspond to input, output, and state variables of a DES model. Thus, the DBN metamodel provides a representation for the joint probability distribution of the input, output, and state variables of the simulation where the state variables are considered at some fixed time instants. The DBNs are used to efficiently calculate marginal and conditional probability distributions of the state variables. The construction and utilization of the DBN metamodels are aided by available BN software (e.g., Decision Systems Laboratory). By using interpolation between the time instants of the DBN, the probability distributions are approximated for any time instants within the duration of the simulation (Poropudas and Virtanen, 2010).

The nodes of DBNs are partitioned into sets corresponding to particular time instants. The sets are called time slices and, typically, all the time slices include nodes corresponding to each of the variables. In the context of DBN metamodels, this means that all state variables are considered at the same time instants. The common time instants are not necessarily ideal because they may ignore changes that are specific to only some variables or, alternatively, include redundant information about others. Such situations arise, e.g., when the changes of one variable occur at a faster pace than others or the changes in the variables take place in distinct time intervals of the simulation.

It is also possible that one variable is considered more important than another for the purposes of the analysis and therefore needs to be treated in more detail.

In this paper, these issues are resolved by utilizing multiple time scales for state variables. In addition, a dynamic programming (DP, e.g., Bertsekas 1995) algorithm similar to (Gluss, 1962) is used to determine the time scales. When multiple time scales are considered in DBN metamodels, the time instants in the DBN are selected independently for each variable. This offers an improvement for the structure of the DBN metamodel. The application of multiple time scales results in a more accurate representation of the time evolution of the simulation without increasing the size of the metamodel.

The paper is organized as follows. The construction of DBN metamodels from simulation data is introduced in Section 2. The utilization of the DBN metamodels in simulation analysis is briefly presented in Section 3. Examples of a DBN metamodel with multiple time scales as well as its application are given in Section 4 where a DBN metamodel is used for probabilistic inference regarding the operation of a simulated air base.

2 Construction of DBN metamodels

The first step in the construction of a DBN metamodel is the selection of variables. While a DES model includes all the variables that significantly affect the behavior of the system, the subset of variables included in the DBN metamodel is selected based on how the DBN is going to be utilized. Assume now that state variables $x_1(t),\ldots,x_n(t)$, where t refers to time, as well as input and output variables $u_1,\ldots,u_m$ and $z_1,\ldots,z_\ell$ of the DES model are included in the metamodel.

The second step in the construction is the design of experiment. Only discrete variables are allowed in DBN metamodels. The values of input variables are therefore discretized, which is discussed in more detail below. When constructing a DBN metamodel, a number of simulation replications are performed for all the combinations of the values of the inputs. A lower limit for the number of replications is calculated based on the objectives of the analyses (for details, see Poropudas and Virtanen 2011). If the number of data is found insufficient later on in the validation step of the construction, additional replications can be performed. The third step of the construction, i.e., the simulation, is performed once the values of the input variables and the number of replications are determined.

Due to the nature of DBNs, the values of state variables x_k are restricted to discrete sets X_k. Thus, the discretization is the fourth step of the construction. The elements of X_k and the manner in which the actual values of x_k are mapped onto them is decided on a case by basis by taking advantage of prior knowledge of the system. If no natural discretization of the variables is available, the values are mapped into a set of discrete bins with the help of general clustering algorithms such as k-means (Hartigan and Wong, 1979). The same procedure is applied to the input and output variables. The input variable u_k obtains values from the discrete set denoted by U_k and the output variable z_k from the discrete set denoted by Z_k.

DBNs are discrete time models where each state variable is considered at a finite number of time instants. In this paper, the time instants are allowed to vary from variable to variable and they are selected separately for each one. This constitutes the fifth step in the construction. The state variable x_k is considered at the time instants

$$T_k = \left\{t_0, t_1^k, t_2^k, \ldots, t_f\right\}, \tag{1}$$

where t_i^k are chosen from the interval (t_0, t_f). Here t_0 and t_f refer to the starting and terminating times of the simulation, respectively, which are assumed to be identical for every replication. The DBN metamodel considers the joint probability distribution of all the variables at all the time instants. The estimate for the probability of the variable x_k obtaining the value $j \in X_k$ at time instant $t \in T_k$, i.e., $P(x_k(t) = j)$, provided by the DBN metamodel is denoted by $\hat{p}_j^k(t)$. A linear interpolation technique is used to construct estimates at the probabilities for time instants that are not included in the DBN. This results in estimates of the form

$$\hat{p}_j^k(t) := \hat{p}_j^k(t_-) + \frac{t - t_-}{t_+ - t_-}\left(\hat{p}_j^k(t_+) - \hat{p}_j^k(t_-)\right), \tag{2}$$

where $t \notin T_k$, $t_- = \max\{v \in T_k | v \le t\}$, and $t_+ = \min\{v \in T_k | v \ge t\}$.

The selection of the time instants T_k begins with the discretization of the time interval $[t_0, t_f]$ into the equally spaced instants

$$T_k^* = \left\{t_0, t_0 + \delta_k, \ldots, t_0 + (m_k - 1)\delta_k, t_f\right\}, \tag{3}$$

where m_k is the number of segments for variable x_k and $\delta_k = (t_f - t_0)/m_k$. The time instants T_k are selected from among the time instants T_k^*. The probability estimates $P(x_k(t) = j) = p_j^k(t)$ of the variable x_k, which are based on the simulation data, are calculated for each time instant T_k^* and each value $j \in X_k$. The time evolution of these probabilities is referred to as the probability curves of the variable x_k. Now, the objective is to select the time instants T_k in such a manner, that the corresponding probabilities provided by the DBN metamodel follow the probability curves closely, while keeping the number of the time instants T_k low.

To quantify the accuracy of the DBN metamodel, the sum of squared error

$$M_k(T_k) = \sum_{t \in T_k^*} \sum_{j \in X_k} \left(p_j^k(t) - \hat{p}_j^k(t)\right)^2, \tag{4}$$

is used. The DBN is constructed so that $p_j^k(t) = \hat{p}_j^k(t)$ for all time instants $t \in T_k$. This means that the probabilities

$\hat{p}_j^k(t)$ are known prior to the construction of the DBN. The problem of selecting the time instants T_k then consists of selecting the number of time instants to include and finding their optimal location which minimizes $M_k(T_k)$.

Since the squared error is summed over all the time instants T_k^*, and the probability estimate given by the DBN for any single time instant depends only on the preceding and the following time instant in T_k, the error is calculated for one segment between consecutive time instants in T_k at a time. The errors are then aggregated to provide the total error $M_k(T_k)$. Thus, it is not necessary to evaluate every potential T_k as a whole because only each pair of time instants needs to be considered separately.

Optimal time instants are found by using dynamic programming in a manner similar to (Gluss, 1962). The algorithm iterates through all the pairs of time instants in T_k^* and calculates the total error for the segment from the one instant $t_0 + a\delta_k$ to another $t_0 + b\delta_k$ by assuming that there are no time instant $t \in T_k$ between them. The total error for such a segment is denoted as

$$D(a,b) = \sum_{i=a}^{b} \sum_{j \in X_k} \Big(p_j^k(t_0 + i\delta_k) - \left(\frac{b-i}{b-a} p_j^k(t_0 + a\delta_k) + \frac{i-a}{b-a} p_j^k(t_0 + b\delta_k) \right) \Big)^2 . \quad (5)$$

The optimization problem is solved by considering subproblems where subsets of the form $\{t_0, t_0 + \delta_k, \ldots, t_0 + b\delta_k\}$ with $0 \leq b \leq m_k$ are considered. Optimal selections of time instants within each such a set are determined. Let $f_l(b)$ denote the resulting minimum error in such a subproblem when l time instants are used to estimate the probability curves between t_0 and $t_0 + b\delta_l$. The time instants t_0 and $t_0 + b\delta_k$ must always be included in the solution, so $l \geq 2$. For all l, $f_l(0) = 0$ and, for all $b \leq m_k$, $f_2(b) = D(0,b)$. For other values of l and b, the value of $f_l(b)$ is determined by the equation

$$f_l(b) = \min_{0 \leq i < b} \{ f_{l-1}(i) + D(i,b) \}, \quad (6)$$

where $D(i,b)$ is given by Eq. (5). $f_l(b)$ is evaluated for each value of l from 2 to the maximum value l_k. Then, the value of b is increased by one and $f_l(b)$ is again evaluated for each value of l. This is repeated until b has gone from 1 to m_k, at which point the algorithm terminates. The optimal time instant sets covering the entire time interval $[t_0, t_f]$ and consisting of any number of time instants up to the maximum l_k have then been calculated. The optimal sets containing different numbers of time instants are compared and the most suitable one is identified.

The structure of the DBN metamodel, consisting of nodes and arcs between them, is determined once T_k is chosen for each variable x_k. The construction is aided by BN software such as GeNIe (Decision Systems Laboratory). In the DBN, nodes are included for each variable x_k at all of the time instants T_k. A node is also associated with each input and output variable. If prior information about the system under consideration is available, the sixth step consists of using this information to define the known dependencies between nodes. Arcs implying dependence between specific nodes can be included regardless of the simulation data. In order to maintain causality, arcs going from a state variable to an input variable, from an output variable to an input variable, from an output variable to a state variable and from a state variable to a state variable at an earlier time instant are not allowed.

The seventh step in the construction consists of finalizing the structure of the network and determining its CPTs. The realized values of each state variable x_k at all of the time instants T_k are recorded for every replication of the simulation model, as are the values of all input and output variables. The structure is completed by applying learning algorithms (Heckerman et al., 1995) on the simulation data. The CPTs are constructed in accordance to the relative frequencies of the values in the data (Poropudas and Virtanen, 2011).

For the input variables, the relative frequencies in the simulation data do not necessarily reflect the actual probability distributions in question because they can be modified as part of the design of experiment step in order to collect a broader set of data. The probability distributions of the input variables are adjusted after the construction of the DBN metamodel in order to represent input certainty (Pousi et al., 2013). The distributions can be modified only after validating the metamodel, because the adjusted distributions for the inputs are not consistent with the validation data.

3 Utilization of DBN metamodels

The constructed DBN metamodel provides the joint probability distribution of the input, output, and state variables. The DBN is applied for various what-if analyses where conditional probabilities and probability distributions are studied. In these analyses, the values of some variables at given time instants are fixed and the conditional probability distributions of the other variables are updated using readily available algorithms implemented in BN software such as GeNIe (Decision Systems Laboratory). When considering conditional probabilities, the chronological order of the time instants is irrelevant, i.e., the conditional probability distributions can be calculated also for conditions related to later time instants.

The most basic application of the DBN metamodel is to determine the marginal probability distribution of a state variable as a function of time. Such a distribution consists of the probabilities of the state variable obtaining a given value at a given time, i.e., $P(x_k(t) = j)$. The marginal probabilities for outputs $P(z_k = j)$ can also be obtained. Conditional probabilities for the state variables are also obtained by setting conditions for inputs $P(x_k(t) = j | u_{k'} = j')$, state variables $P(x_k(t) = j | x_{k'}(t') = j')$, outputs $P(x_k(t) = j | z_{k'} = j')$, or any combination of

these. Conditions can even be set for the same state variable that is being investigated, as long as the time instants are different. Conditional probabilities for output variables are calculated similarly. To create input-output mappings, conditions are set for input variables and the conditional probability distributions of the outputs, e.g., $P(z_k = j|u_{k'} = j')$, are studied. Conditional distributions of the outputs can also be studied by fixing the values of state variables, e.g., $P(z_k = j|x_{k'}(t) = j')$. If the metamodel includes multiple output variables, conditions can be set for some of them as well resulting in conditional probabilities such as $P(z_k = j|z_{k'} = j')$. The conditional probability distributions for input variables are investigated by setting conditions for the state variables $P(u_k = j|x_{k'}(t) = j')$, output variables $P(u_k = j|z_{k'} = j')$, or both. This inverse reasoning can be used to investigate, e.g., which combination of input values is most likely to lead to a certain outcome.

If the analysis involves probability distributions related to time instants not included in the DBN, the interpolation discussed in Section 2 is applied. The interpolation can also be applied to conditions taking place at time instants not included in the DBN. The details of the interpolation scheme are presented in (Poropudas and Virtanen, 2010).

4 Example analysis - simulated operation of air base

In this example, a DBN metamodel is constructed according to the guidelines discussed in Section 2 and used in the simulation analysis of the operation of an air base. In the model, aircraft go a through a cycle consisting of mission assignment, mission execution, repair of possible damage obtained during the mission, and standard service such as fueling. There are three queues for the aircraft: one for mission assignment, one for repair, and one for service. The repair and service personnel can only work on one aircraft at a time. The aircraft that have not been damaged move directly from the mission to the service queue. An aircraft is released from the mission assignment queue every time a new mission is to be executed. If there are no aircraft in this queue, a backlog of missions is formed and the aircraft are assigned to the missions as soon as they arrive from the service. A flowchart of the simulation model is presented in Fig. 1.

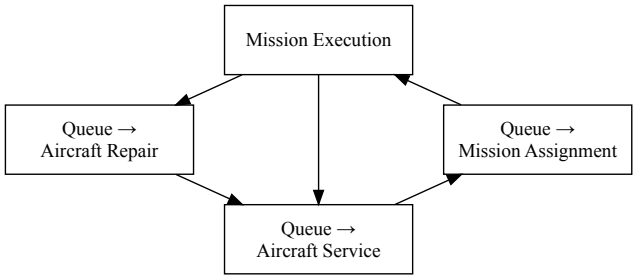


Figure 1. Flowchart of the simulation model.

The missions are categorized into patrol missions and combat missions. The patrol missions are assigned regularly with the time between consecutive missions sampled from a uniform probability distribution. The aircraft are unlikely to be damaged during a patrol mission. Combat missions are assigned as a Poisson process with a time dependent arrival intensity. They are on average shorter than the patrol missions but the aircraft have a much higher probability of being damaged. The repair time of a damaged aircraft is exponentially distributed. The service time is deterministic and depends on the length and type of the preceding mission.

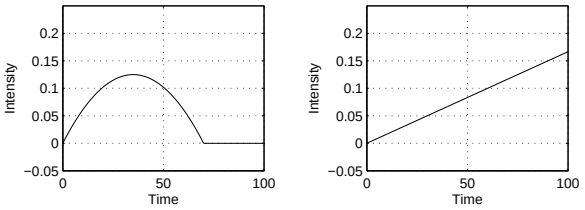
Table 1. Variables of the metamodel.

	Type	Range	Interpretation
u	Input	$\{1,\ldots,4\}$	Scenario
x_1	State	$\{0,\ldots,4\}$	Aircraft in assignment
x_2	State	$\{0,\ldots,4\}$	Aircraft in repair
z	Output	$\{0,1\}$	Insufficient aircraft available

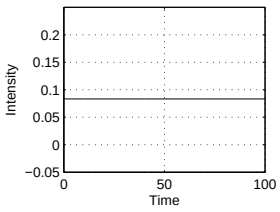
The input variable of the simulation model, denoted by u, determines the time dependent intensity of the occurrence of the combat missions. In this example, four alternative scenarios are studied. The number of aircraft in each of the four locations are considered as state variables. Two of the state variables, i.e., the number of aircraft in mission assignment, denoted by $x_1(t)$, and in aircraft repair, denoted by $x_2(t)$, are included in the DBN metamodel. The output variable of the simulation model is an indicator, denoted by z, that determines whether or not at any time during the simulation no aircraft are available to execute an incoming mission. The variables included in the DBN and their ranges are summarized in Table 1.

In order to acquire data for construction of the DBN, four scenarios are simulated. In the first one, the arrival intensity of the combat missions starts at 0, peaks early in the simulation, and returns to 0 later on. In the second scenario, the intensity slowly increases throughout the simulation. In the third one, the intensity is constant. In the fourth scenario, there are no combat missions. The non-zero arrival intensities of the first three scenarios are illustrated in Fig. 2. The four scenarios occur with equal probability. In every simulation replication, four aircraft are included. The data is collected by running 2000 simulation replications for each scenario. The duration of each replication is 100 units of time. A quarter of the data is reserved for validation. Since all the variables under consideration are discrete, there is no need for their discretization.

The probability curves of the state variables calculated from the simulation data are shown in Fig. 3. The advantage of utilizing multiple time scales is evident. The probability distribution of x_1 changes repeatedly due to the regularly scheduled patrol missions while the distribution of x_2 changes more slowly. Determining the optimal time instants for the state variables using the DP algorithm discussed in Section 2, a suitable number of time instants for

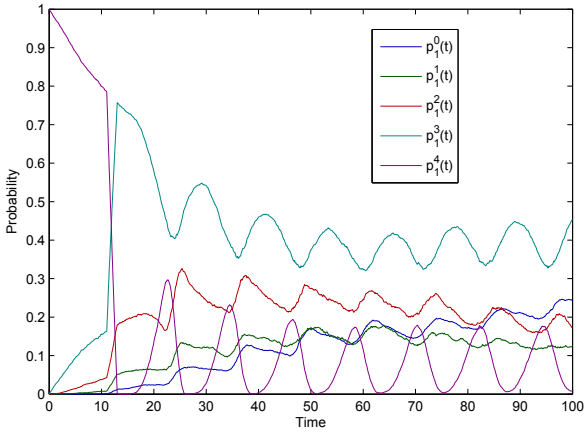


(a) Scenario $u = 1$. **(b)** Scenario $u = 2$.

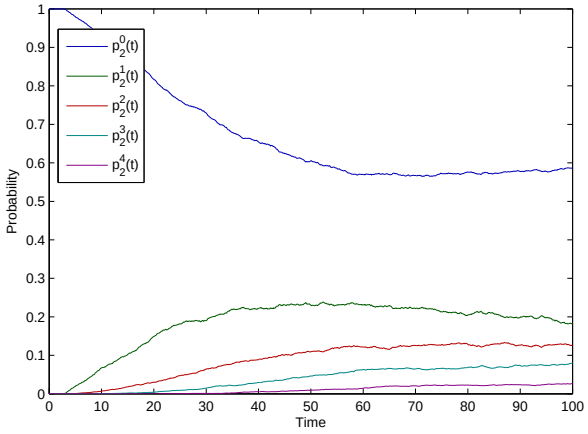


(c) Scenario $u = 3$.

Figure 2. Intensity of the generation of combat mission in three scenarios.



(a) Number of aircraft in mission assignment, x_1.



(b) Number of aircraft in aircraft repair, x_2.

Figure 3. Time evolution of the marginal probability distribution of the state variables as estimated from the simulation data.

x_1 is found to be 25. For x_2, seven time instants are used.

Prior knowledge is used to add arcs to each node in the DBN from the most recent node corresponding to each variable. Arcs originating from nodes corresponding to the time instant 0, except those leading to the following node corresponding to the same variable, are ignored, since the initial state of the simulation is always same. Arcs are also added from the input variable to all other nodes and from every other node to the output variable. The arcs determined in this manner are sufficient to describe the entire system because no additional dependencies are evident in the simulation data.

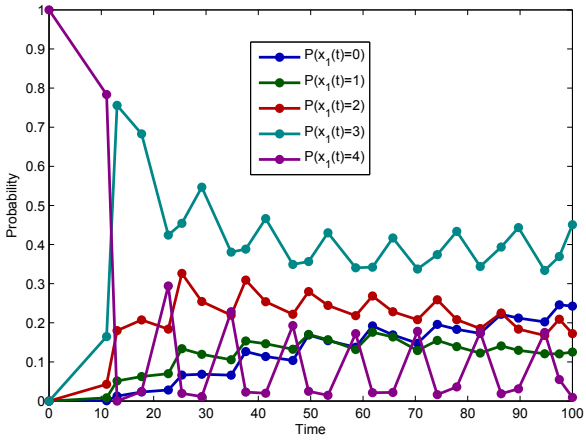
Fig. 4 depicts the unconditional time evolution of the simulation, i.e., the time evolution of the marginal probability distributions of the two state variables provided by the DBN metamodel. The distributions resemble the probability curves in Fig. 3. The periodical nature of x_1, caused by the regular patrol missions, is evident in Fig. 4a. This is also the main reason why the concept of multiple time scales is useful in this example. The patrol missions directly affect the number of aircraft available for missions, but have little impact on the number of aircraft needing repair.

Next, alternative what-if analyses allowed by the DBN metamodel are illustrated. The first of the four scenarios is examined in Fig. 5 by setting the condition $u = 1$. When comparing to Fig. 4, there are fewer aircraft ready for missions and more in need of repair during the middle of the time interval but the situation is reversed by the end of it. This is consistent with the intensity of the generation of combat missions presented in Fig. 2.

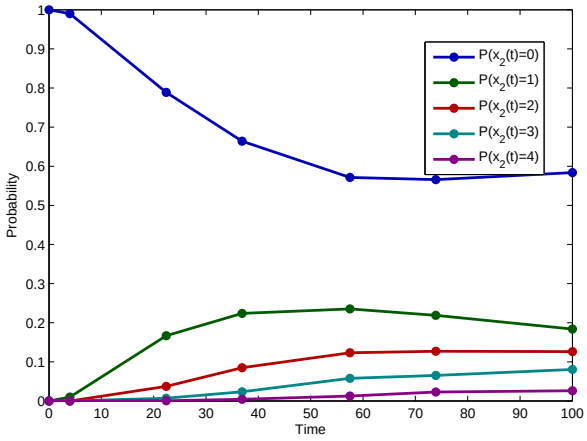
In order to further investigate conditional properties of the simulation model, the condition regarding u is removed and the condition $x_1(100) = 0$ is instead added, i.e., every aircraft is either on a mission, being repaired, being serviced or in a queue waiting for one of the latter two activities at the end of the simulation. The probability of this event is $P(x_1(100) = 0) = 0.24$. Fig. 6 presents the time evolution of the conditional probability distributions of the state variables. The likely number of aircraft ready for mission decreases steadily in Fig. 6a. The expected number of aircraft in need of repair increases conversely in Fig. 6b, but the most likely values of $x_2(100)$ are 2 and 3 which means that one or two aircraft are probably still either carrying out a mission, being serviced, or waiting for service.

The condition $x_1(100) = 0$ also affects the output variable z. The probability distribution of z without and with the condition is presented in Table 2. The condition greatly increases the probability of z obtaining the value 1. This is as expected since a mission with no aircraft available to carry it out can only occur if x_1 obtains the value 0 at some point.

This example demonstrates just some of the capabilities of DBN metamodels with multiple time scales. With more variables and replications of the simulation model, more elaborate what-if analyses can be performed. By fully utilizing existing BN software, this can be done programmat-

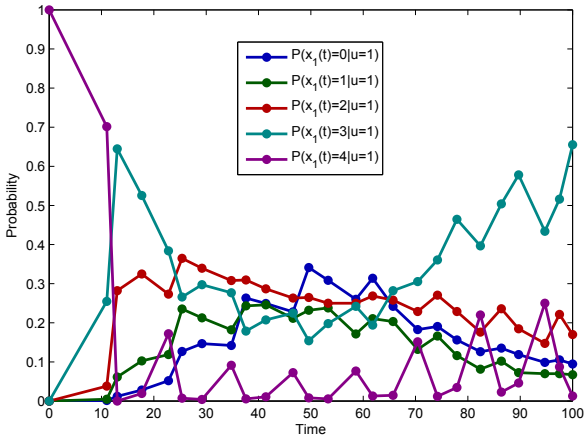


(a) Number of aircraft in mission assignment, x_1.

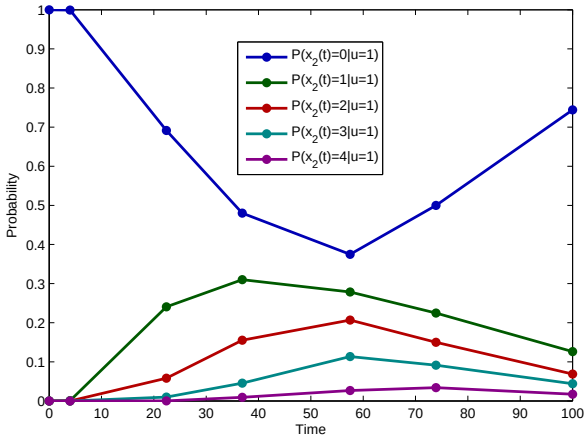


(b) Number of aircraft in aircraft repair, x_2.

Figure 4. Time evolution of the marginal probability distributions of the state variables provided by the DBN metamodel.



(a) Number of aircraft in mission assignment, x_1.



(b) Number of aircraft in aircraft repair, x_2.

Figure 5. Time evolution of the conditional probability distributions of the state variables provided by the DBN metamodel conditional on $u = 1$.

ically, which greatly enhances the number of probability estimates that can be calculated in a reasonable amount of time.

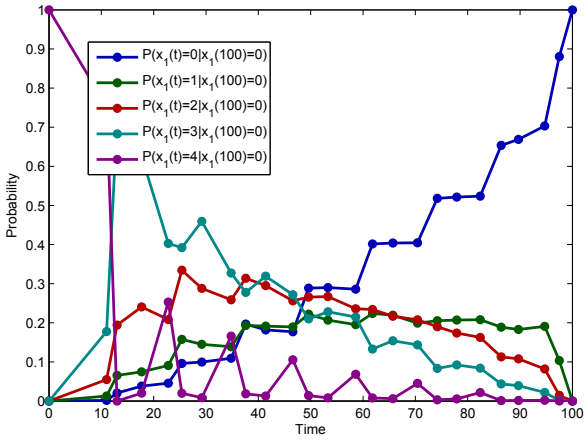
5 Conclusions

The time evolution of DES can be analyzed in a transparent manner by using DBNs as simulation metamodels. The DBN metamodels offer an effective way for conducting various what-if analyses. In the previous literature, the structure of DBN metamodels has consisted of time slices, i.e., the networks have had a rigid structure where all state variables of the model are considered at each of the time instants represented by the DBN. In this paper, the concept of multiple time scales is introduced to the DBN metamodeling, i.e., the changes in the probability distributions of the state variables are allowed to occur independently in the DBN and the time evolution of individual state variables is studied at its own pace. By employing multiple time scales, the different temporal changes in the behavior of the state variables are described more accurately without unnecessary increase in the size of the DBN. The paper also presents an algorithm based on dynamic programming for the optimal selection of time instants represented by the DBN. The construction and utilization of the DBN metamodel with multiple time scales are demonstrated with an example analysis involving the operation of an air base.

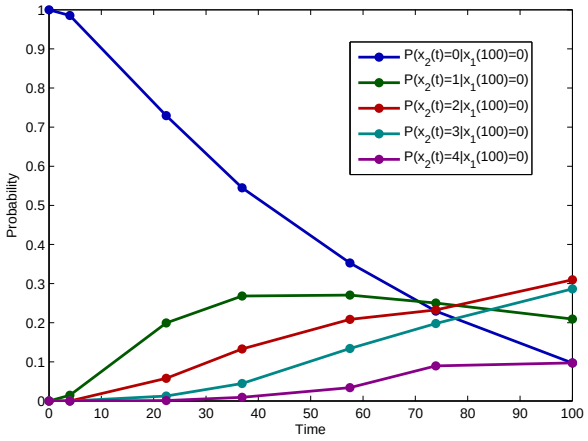
Simulation studies using DBN metamodels can be performed with software designed for the analysis of BNs. Unfortunately, the dynamic programming algorithm and the interpolation technique used for approximative reasoning are beyond the scope of such software and, thus, the calculations presented in this paper have been carried out using MATLAB. In order to alleviate future studies, it is worthwhile to develop an automated tool designed for the construction and utilization of DBN metamodels.

Table 2. Marginal and conditional probability distributions of the output variable when $x_1(100) = 0$.

j	$P(z = j)$	$P(z = j \mid x_1(100) = 0)$
0	0.63	0.17
1	0.37	0.83



(a) Number of aircraft in mission assignment, x_1.



(b) Number of aircraft in aircraft repair, x_2.

Figure 6. Time evolution of the conditional probability distributions of the state variables provided by the DBN metamodel when $x_1(100) = 0$.

The DBN metamodels have also been used in simulation-based optimization as a part of influence diagram metamodels (Poropudas and Virtanen, 2009). In such metamodels, the DBN reveals the consequences of decision alternatives, i.e., the time evolution of a simulated system with given values of simulation parameters. Clearly, the concept of multiple time scales could also be applied in the construction of influence diagram metamodels from simulation data.

References

D.P. Bertsekas. *Dynamic programming and optimal control*. Athena Scientific Belmont, MA, 1995.

Decision Systems Laboratory. GeNIe (graphical network interface). Available via https://www.bayesfusion.com/ [accessed November 8, 2017].

L.W. Friedman. *The simulation metamodel*. Springer Science & Business Media, 2012.

B. Gluss. Further remarks on line segment curve-fitting using dynamic programming. *Communications of the ACM*, 5(8): 441–443, 1962.

J.A. Hartigan and M.A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

J.P.C. Kleijnen. *Design and analysis of simulation experiments*, volume 20. Springer, 2008.

A.M. Law and W.D. Kelton. *Simulation modeling and analysis*. McGraw Hill Boston, 2000.

K.P. Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.

J. Pearl. Fusion, propagation, and structuring in belief networks. *Artificial intelligence*, 29(3):241–288, 1986.

J. Poropudas. *Bayesian Networks, Influence Diagrams, and Games in Simulation Metamodeling*. Doctoral dissertation, Aalto University School of Science, 2011. Available via http://lib.tkk.fi/Diss/2011/isbn9789526042688/ [accessed November 8, 2017].

J. Poropudas and K. Virtanen. Analyzing air combat simulation results with dynamic Bayesian networks. In *Proceedings of the 2007 Winter Simulation Conference*, pages 1370–1377. IEEE Press, 2007.

J. Poropudas and K. Virtanen. Influence diagrams in analysis of discrete event simulation data. In *Proceedings of the 2009 Winter Simulation Conference*, pages 696–708. Winter Simulation Conference, 2009.

J. Poropudas and K. Virtanen. Simulation metamodeling in continuous time using dynamic bayesian networks. In *Proceedings of the 2010 Winter Simulation Conference*, pages 935–946. Winter Simulation Conference, 2010.

J. Poropudas and K. Virtanen. Simulation metamodeling with dynamic Bayesian networks. *European Journal of Operational Research*, 214(3):644–655, 2011.

J. Pousi, J. Poropudas, and K. Virtanen. Simulation metamodelling with Bayesian networks. *Journal of Simulation*, 7(4): 297–311, 2013.

B.P. Zeigler, H. Praehofer, and T.G. Kim. *Theory of modeling and simulation: integrating discrete event and continuous complex dynamic systems*. Academic press, 2000.

Size Rate of an Alternative Aggregation Petri net developed under a Modular Approach

Juan-Ignacio Latorre-Biel [1] Emilio Jiménez-Macías [2] Julio Blanco [2] Mercedes Perez [2]

[1] Department of Mechanical Energetic and Materials Engineering, Public University of Navarre, Spain,
juanignacio.latorre@unavarra.es

[2] High Technical School of Industrial Engineering, University of La Rioja, Spain,
{emilio.jimenez,julio.blanco,mercedes.perez}@unirioja.es

Abstract

Petri nets allow describing formal models of discrete event systems, which might show counterintuitive behaviors. The design of a discrete event system, composed by known subsystems, requires the definition of the interrelations between them. This feature can be modeled in the structure of the Petri net by arcs and link transitions. The choice of the best configuration might be a hard problem to solve due to the foreseeable combinatorial explosion. In order to alleviate the computer resources required for exploring the different feasible combinations of the subnets, a single model with exclusive entities can be developed by an alternatives aggregation Petri net. In this paper the construction of such a model with four subnets and certain precedence constrain is discussed. Also, a reduction in the size of the amount of required information for describing the alternative structural configurations is calculated for different sizes of the subnets.

Keywords: Petri nets, modelling and simulation, modular design, alternative structural configuration, decision support system

1 Introduction

Petri nets constitute a paradigm in the modeling of discrete event systems (Silva, 1993). The simplicity of its rules, the double representation of a model, both graphical and matrix-based, as well as its ability to describe features, such as parallelism, precedence, concurrence, synchronization, or competence for shared resources, makes Petri nets an invaluable tool for applications such as performance evaluation or structural analysis (David and Alla, 2005).

One fruitful application of Petri nets is the field of decision making support (Latorre *et al*, 2014c). However, other methodologies can be considered for this task (Bruzzone and Longo, 2010). Among them, modeling and simulation have been applied successfully (Jiménez-Macías and Pérez-Parte, 2004; Piera *et al*, 2004; Longo *et al*, 2013; Mújica *et al*, 2010). In particular, Petri nets can be used for quasi-optimal operation or design of complex systems (Latorre *et al*, 2014b). In particular, a design process usually requires choosing between different alternative structural configurations, which makes this kind of decision problem singular in nature (Latorre and Jiménez, 2013).

Some particular problems of design are tackled by the selection of certain subsystems, such as particular machines, manufacturing lines, or even manufacturing facilities, and the ulterior choice of the way, these systems are related. These decisions involve the definition of the interchange of information, parts, products, vehicles, persons, or whatever the flow between subsystems is. Moreover, these decisions configure the behavior of the subsystems by features such as precedence, synchronization, or parallelism (Latorre *et al*, 2014a).

A classic methodology to address this decision problem starts considering a different model, or alternative Petri net, for every alternative structural configuration of the system (Latorre *et al*, 2014c). This approach, however, presents some drawbacks, such as the need to develop a large number of models and to analyze specifically every one of them, or to discard good decisions by reasons, such as intuition, rough analysis, personal preferences, lack of awareness, etc. Another important drawback is the large amount of data required for representing all the alternative Petri nets, since they may be created by different combinations of the shared subnets (Latorre-Biel *et al*, 2015).

This combinatorial process for constructing feasible solutions of the design problem, implies the fact that many data required to represent a set of alternative Petri nets is redundant: every shared Petri net belongs to many models and its description is repeated every time, increasing in this way the size of the description of the system with alternative structural configurations (w/ASC) (Latorre *et al*, 2014b).

Moreover, this is not the only type of redundant information, present in a set of alternative Petri nets, since a given transition between subnets or link transitions, can also be present in several alternative Petri nets (Latorre-Biel *et al*, 2015).

In order to overcome these drawbacks, a family of Petri net-based formalisms has been developed. All of

them can model a discrete event system w/ASC. The exclusiveness of the alternative structural configurations is represented by a set of exclusive entities. In particular, the alternatives aggregation Petri nets can integrate in a natural way a set of shared subnets (Latorre *et al*, 2013).

This paper deals with the design process of a discrete event system, once a set of subsystems has been chosen. Another constrain of the problem is a relation of strict precedence of one of the subnets, and the other three.

The rest of the paper is organized as follows. Section 2 defines the formalism that will be considered for developing the model of the system: the AAPN. Section 3 states in a detailed way the design problem to be solved. The three following sections (Section 4, Section 5 and Section 6) present different types of relations between the subsystems and provide with the expressions to calculate the size rate of the AAPN. Section 7 addresses the complete AAPN model, representing the cases discussed in the previous sections, compares the size rates, and comments the trend of the size rate of all the cases as the size of the shared subnets grow. Finally, the conclusions derived from this piece of research are stated in Section 8.

2 Alternatives Aggregation Petri Nets

An AAPN is a formalism that contains a set of exclusive entities. On the other hand, a set of exclusive entities is a collection of mathematical elements representing the exclusiveness that characterizes the alternative structural configurations of the modeled system, i.e. only one of them can be chosen as a result of a decision. In particular, the exclusive entities in an AAPN are the so called choice variables, Boolean variables that configure the guard functions of certain link transitions in the model.

An AAPN is defined by a set of subnets, some of which are shared by different alternative Petri nets, another set of link transitions between the subnets, and a last set of guard functions of choice variables associated to some link transitions.

Alternatives aggregation Petri nets (AAPN) can be applied successfully in the modeling process of discrete event systems with alternative structural configurations (w/ASC), where the different configurations present common or shared subnets.

In this case, the number of redundant information in the form of shared subnets that can be removed in the AAPN may be significant. Shared subnets contain places and transitions and their removal contribute to a reduction in the size of the incidence matrix of the AAPN, resulting, for example, in a speed up of the simulation of the net, useful for performance analysis.

The construction process of an AAPN from a set of alternative Petri net is quite straightforward. The following steps can be followed to achieve this objective:

a) Decompose the alternative Petri nets into subnets and link transitions. The election of the limits of every subnet is a choice of the modeler. However, certain criteria can be considered, such as associating each subnet with a physical element of the real system or trying to guarantee that every subnet is shared by the largest number of alternative Petri nets.

b) Take one of the alternative Petri nets as seed of the AAPN. Associate the first choice variable to every link transition.

c) Consider the following alternative Petri net (ith alternative Petri net) and compare it to the seed of the AAPN. Every subnet of the alternative Petri net that does not belong to the AAPN should be added to the AAPN. Add to the AAPN all the link transtions of the alternative Petri net. Every added link transition should be associated to the ith choice variable.

d) Apply reduction rules to the quasi-identical link transitions of the AAPN.

e) Repeat steps c) and d) until all the alternative Petri nets have been added to the seed of the AAPN and the final AAPN model is complete.

With these considerations as background, the following section will define the scope of the design problem to be solved.

3 Statement of the problem

The objective of this paper is to show the feasibility, the methodology, and some advantages of constructing an AAPN for decision making support in the design process of a kind of discrete event system.

In particular, the system to be designed should include a single unit of every one of four different subnets $\{R_A, R_B, R_C, R_D\}$. These subnets are chosen to have only an input and an output link transitions. In addition, the input transition presents a single output place. Analogously, the output transition has a single input place.

Moreover, one of the subnets should comply with a relation of strict precedence with the other three subnets, which should evolve in parallel, simultaneously and/or alternatively.

Moreover, solutions with other precedence relations between the subnets, leading to an unbounded Petri net, or presenting deadlocks, should be discarded.

Complying with the mentioned constraints, three options with different types of relations between the subnets are considered and analyzed in the following sections.

The following notation will be considered:

A_r, B_r, C_r, D_r, are the number of rows of the incidence matrix of the alternative Petri nets R_A, R_B, R_C, R_D respectively.

A_c, B_c, C_c, D_c, are the number of columns of the incidence matrix of the alternative Petri nets R_A, R_B, R_C, R_D respectively.

4 Case I

The first structure for the solution of the decision problem has been drawn in Figure 1, where a general structure of an alternative Petri net has been represented. In this figure, the four different subnets are depicted by means of clouds, which is an informal or incomplete description of a Petri net, since the input and output places of the link transitions, belonging to every subnet, are not specified.

Furthermore, the internal structure of the subnet is not detailed, due to the fact that the purpose of this representation is to develop a general description, where the relationship between the subnets is pointed out. The link transitions are depicted explicitly in the representation of Figure 1.

In this figure, the different subnets are represented by a general name of R_w, R_x, R_y, R_z, since all the combinations of the shared subnets $\{R_A, R_B, R_C, R_D\}$ can be substituted in the positions of $\{R_w, R_x, R_y, R_z\}$ defining feasible solutions for the discrete event system to be designed. In fact, there are four feasible combinations of subnets in this case 1, leading to four alternative Petri nets $\{R_1, R_2, R_3, R_4\}$. Two of them are depicted in Figure 2.

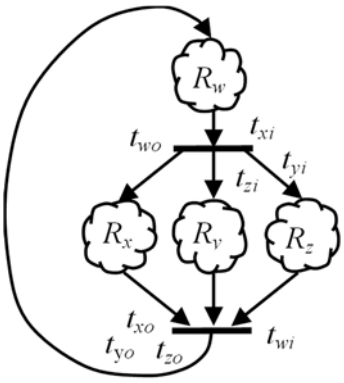


Figure 1. Structure of the link transitions for the case 1 as solution of the discrete event system to be designed.

It can be seen that the link transitions can be named after their input or output subnets. Since the four subnets are connected to each link transition, any of the link transitions can have four different names, as it can be seen in Figure 2.

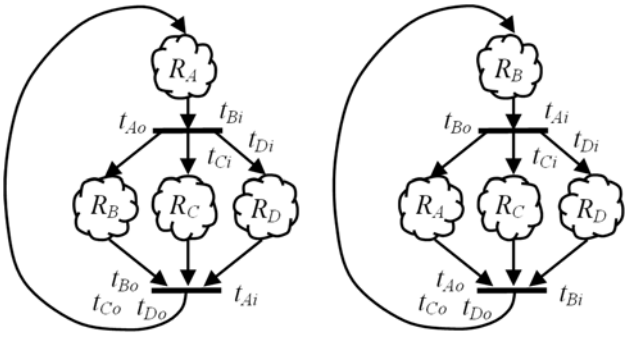


Figure 1. Two feasible solutions of case I: R_{1a} and R_{2a}.

Following the steps detailed in section 2, it is possible to construct a single alternatives aggregation Petri net from the complete set of four alternative Petri nets. This AAPN contains the four subnets $\{R_A, R_B, R_C, R_D\}$, as well as two link transitions from every alternative Petri net $\{R_{1a}, R_{2a}, R_{3a}, R_{4a}\}$. Due to the fact that it is not possible to find quasi-identical transitions among the link transitions of the AAPN, it is not possible to apply a reduction rule to diminish the number of link transitions. For this reason, the number of link transitions of the AAPN is $2 \cdot 4 = 8$.

Let us consider the following notation:

r and c are the number of rows and columns of an alternative Petri net respectively.

r' and c' are the number of rows and columns of the resulting AAPN respectively.

It has to be considered that the size of the four alternative Petri nets $\{R_{1a}, R_{2a}, R_{3a}, R_{4a}\}$ is the same and can be calculated as follows:

$$r = A_r + B_r + C_r + D_r$$
$$c = A_c + B_c + C_c + D_c + 2 \quad (1)$$

where the number 2 added to the calculation of c comes from the two link transitions of every alternative Petri net $\{R_{1a}, R_{2a}, R_{3a}, R_{4a}\}$.

Moreover, the size of the incidence matrix of the AAPN is:

$$r' = A_r + B_r + C_r + D_r$$
$$c' = A_c + B_c + C_c + D_c + 8 \quad (2)$$

Let us call $x = A_c + B_c + C_c + D_c$; hence,

$$c = x + 2$$
$$c' = x + 8 \quad (3)$$

It is possible to calculate the reduction size of the AAPN, when compared with the set of alternative Petri nets $S_R = \{R_{1a}, R_{2a}, R_{3a}, R_{4a}\}$. Both of them, the AAPN and S_R represent the same system and contain exactly the same information. However, the amount of data required by any of them is quite different:

$$\text{size rate} = \text{size(AAPN)} / \text{size}(S_R) = r' \cdot c' / (4 \cdot r \cdot c) \quad (4)$$

According to (1) and (2), it is possible to state that

$$r = r' = A_r + B_r + C_r + D_r \quad (5)$$

as a consequence

$$\text{size rate} = c' / (4 \cdot c) \quad (6)$$

Moreover, considering (3) it is obtained that

$$\text{size rate} = (x+8) / [4 \cdot (x+2)] \quad (7)$$

Figure 3, represents the trend of the size rate for different values of $x = A_c + B_c + C_c + D_c$, which is the addition of the number of places belonging to every shared subnet. In particular, it can be seen that as x increases, the size rate decreases to a limit value given by:

$$\text{size rate} = (1+8/x) / [4 \cdot (1+2/x)] \quad (8)$$

Calculating the limit of the previous expression as x increases to infinity is $1/4 = 0.25$.

This means that as the number of places in the shared subnets grows, the size of the AAPN approaches to 25%

of the size of the original set of alternative PN S_R. In other words, with a 25% of the data, the AAPN provides the same modeling information than the set S_R.

The axis of abscissas in Figure 3 has been represented with a logarithmic scale with the purpose of detailing the trend for small sizes of the shared subnets.

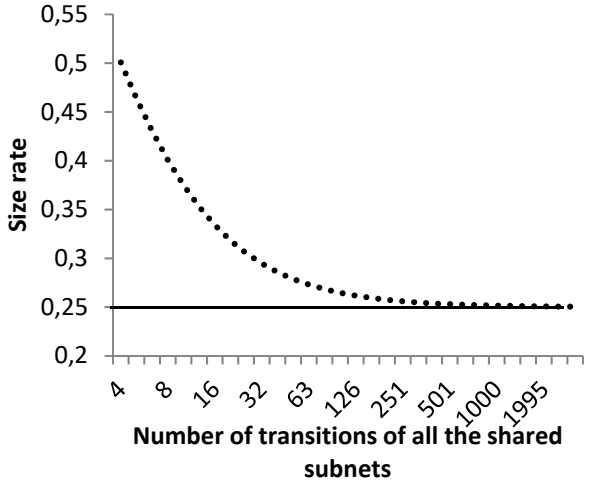


Figure 3. Size rate of and AAPN representing a complete set of alternative Petri nets in case I.

5 Case II

This second case corresponds to another structure for the link transitions in the construction of alternative Petri nets that verify the constraints of the problem stated in section 3. Figure 4 depicts two examples of the four feasible combinations of the shared subnets for constructing such alternative Petri nets.

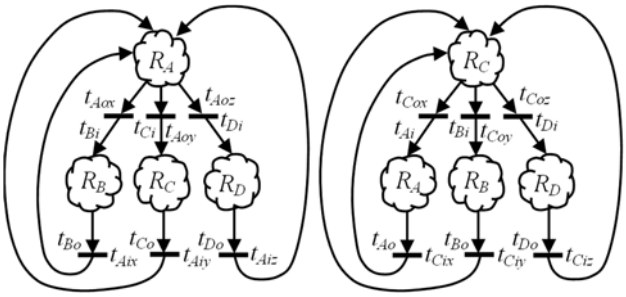


Figure 4. Two feasible solutions of case II: R_1 (left) and R_3 (right).

The four subnets $\{R_A, R_B, R_C, R_D\}$ can be combined in four different ways for constructing four alternative Petri nets $S_{Rb} = \{R_{1b}, R_{2b}, R_{3b}, R_{4b}\}$. Every one of these alternative Petri nets presents 6 link transitions. Moreover, the application of the steps for constructing an AAPN, described in section 2, allow obtaining a single Petri net, representing the complete set of alternative Petri nets S_{Rb}.

The resulting AAPN presents 24 link transitions. However, every link transition has a quasi-identical transition in this set. It is possible to find a couple of examples in Figure 4. In particular, transition t_{Ci} of R_1 is quasi-identical to t_{Cix} in R_3. Moreover t_{Co} in R_1 is quasi-identical to t_{Cox} in R_3. In fact, these transitions are not identical due to the fact that in the AAPN, a different choice variable is associated to each transition from every couple of quasi-identical transitions.

As a consequence of the previous considerations 12 quasi-identical transitions of the AAPN can be combined with their quasi-identical counterparts, leading a Petri net with only 12 link transitions from the original 24 (6 from every original alternative Petri net).

It has to be considered that the size of incidence matrices of the four alternative Petri nets $\{R_{1b}, R_{2b}, R_{3b}, R_{4b}\}$ is the same, $r \cdot c$, where:

$$r = A_r + B_r + C_r + D_r$$
$$c = A_c + B_c + C_c + D_c + 6 \quad (9)$$

where 6 is the number of link transitions of any of the original alternative Petri nets $\{R_{1b}, R_{2b}, R_{3b}, R_{4b}\}$.

Analogously, the size of the incidence matrix of the AAPN is $r' \cdot c'$, where:

$$r' = A_r + B_r + C_r + D_r$$
$$c' = A_c + B_c + C_c + D_c + 12 \quad (10)$$

Let us call $x = A_c + B_c + C_c + D_c$; hence,

$$c = x + 6$$
$$c' = x + 12 \quad (11)$$

The reduction size of the AAPN is:

$$\text{size rate} = \text{size (AAPN)} / \text{size } (S_{Rb}) = r' \cdot c' / (4 \cdot r \cdot c)$$

According to (9) and (10), it is possible to state that

$$r = r' = A_r + B_r + C_r + D_r \quad (12)$$

as a consequence

$$\text{size rate} = c' / (4 \cdot c) \quad (13)$$

Moreover, considering (11) it is obtained that

$$\text{size rate} = (x+12) / [4 \cdot (x+6)] \quad (14)$$

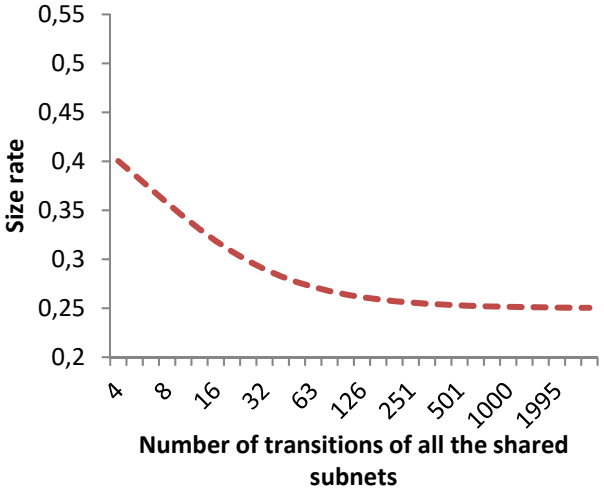


Figure 5. Size rate of and AAPN representing a complete set of alternative Petri nets in case II.

Figure 5 represents the trend of the size rate for different values of $x = A_c + B_c + C_c + D_c$, which is the addition of the number of places belonging to every shared subnet. In particular, it can be seen that as x increases, the size rate decreases to a limit value given by:

$$\text{size rate} = (1+12/x) / [4 \cdot (1+6/x)] \quad (15)$$

Calculating the limit of the previous expression as x increases to infinity is 1/4 = 0.25, the same as in case I. However, the curve that represents the size rate in case I is different to the one representing the same parameter in case II due to a different function of x: (7) versus (14). In fact, the size rate corresponding to case II is smaller than the size rate of case I.

Figure 5 represents the trend of the size rate of the AAPN obtained for case II.

6 Case III

A different structure for the link transitions in the construction of alternative Petri nets that verify the constraints of the problem stated in section 3 is presented in this case III. Figure 6 represents two examples of the twelve feasible combinations of the shared subnets for constructing such alternative Petri nets.

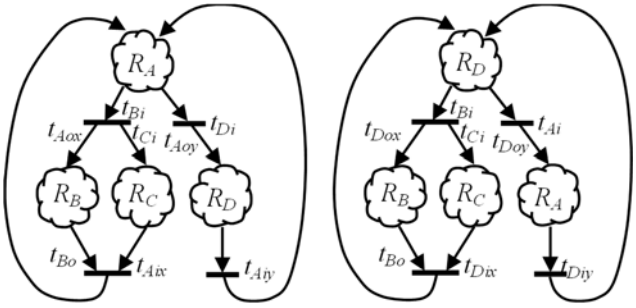


Figure 6. Two feasible solutions of case III: R_1 (left) and R_{12} (right).

In case III The four subnets $\{R_A, R_B, R_C, R_D\}$ can be combined in twelve different ways for constructing twelve alternative Petri nets $S_{Rc} = \{R_{1c}, R_{2c}, \dots, R_{12c}\}$. Every alternative Petri net of S_{Rc} contains 4 link transitions. Furthermore, it is possible to apply to S_{Rc} the steps mentioned in section 2 for obtaining and equivalent AAPN.

As a result, an AAPN with 48 link transitions can be obtained. In this Petro net 12 couples of quasi-identical transitions can be found. Just to give two examples that have been depicted in Figure 6, it is possible to consider transition t_{Di} of R_1 is quasi-identical to t_{Diy} in R_{12}. Moreover t_{Aiy} in R_1 is quasi-identical to t_{Ai} in R_{12}.

As a consequence of the previous considerations 12 quasi-identical transitions of the AAPN can be combined with their quasi-identical counterparts, leading a Petri net with only 36 link transitions from the original 48 (4 from every original alternative Petri net).

The incidence matrices' size of $S_{Rc} = \{R_{1c}, R_{2c}, \dots, R_{12c}\}$ is:

$$r = A_r + B_r + C_r + D_r$$
$$c = A_c + B_c + C_c + D_c + 4 \quad (16)$$

where 4 is the number of link transitions of any of the original alternative Petri nets $S_{Rc} = \{R_{1c}, R_{2c}, \dots, R_{12c}\}$.

Analogously, the size of the incidence matrix of the AAPN is $r' \cdot c'$, where:

$$r' = A_r + B_r + C_r + D_r$$
$$c' = A_c + B_c + C_c + D_c + 36 \quad (17)$$

Let us call $x = A_c + B_c + C_c + D_c$; hence,

$$c = x + 4$$
$$c' = x + 36 \quad (18)$$

The reduction size of the AAPN is:

$$\text{size rate} = \text{size (AAPN)} / \text{size } (S_{Rb}) = r' \cdot c' / (4 \cdot r \cdot c) \quad (19)$$

According to (16) and (17), it is possible to state that

$$r = r' = A_r + B_r + C_r + D_r \quad (20)$$

as a consequence

$$\text{size rate} = c' / (12 \cdot c) \quad (21)$$

Moreover, considering (18) it is obtained that

$$\text{size rate} = (x+36) / [12 \cdot (x+4)] \quad (22)$$

Figure 7, represents the trend of the size rate for different values of x. In particular, it can be seen that as x increases, the size rate decreases to a limit value given by

$$\text{size rate} = (1+36/x) / [12 \cdot (1+4/x)] \quad (23)$$

Calculating the limit of the previous expression as x increases to infinity is 1/12 = 0.0833.

Figure 7 represents the trend of the size rate of the AAPN obtained for case III.

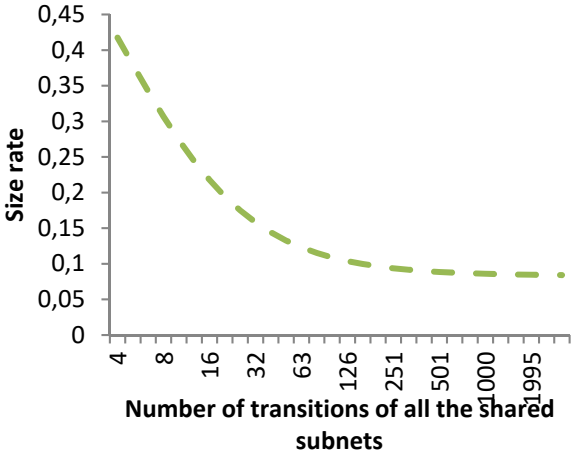


Figure 7. Size rate of and AAPN representing a complete set of alternative Petri nets in case III.

7 Complete model

The steps for constructing an AAPN can also be applied to all the alternative Petri nets in the sets S_{Ra}, S_{Rb}, and S_{Rc}, defined in the previous sections.

As a result, a single AAPN can be developed to represent the 20 alternative Petri nets defined by the sets S_{Ra}, S_{Rb}, and S_{Rc}. The size rate of this AAPN, compared to the original sets of alternative Petri nets can be defined as follows:

$$\text{size rate} = \text{size (AAPN)} / [\text{size } (S_{Ra}) + \text{size } (S_{Rb}) + \text{size } (S_{Rc})] = (r' \cdot c') / [\,4 \cdot r \cdot (x+2) + 4 \cdot r \cdot (x+6) + 12 \cdot r \cdot (x+4)\,] \quad (24)$$

where $r' = r$; hence,

$$\text{size rate} = c' / [\,4 \cdot (x+2) + 4 \cdot (x+6) + 12 \cdot (x+4)]$$
$$c' = x + 8 + 12 + 36 - 12 \quad (25)$$

In this last expression, the added numbers correspond to the link transitions provided to the global AAPN by the partial AAPN constructed from S_{Ra}, S_{Rb}, and S_{Rc} respectively. The negative number corresponds to quasi-identical transitions. As a result:

$$\text{size rate} = (x + 44) / (20 \cdot x + 80) \quad (26)$$

Calculating the limit of the previous expression as x increases to infinity is 1/20 = 0.05.

Figure 8 represents the trend of the size rate of the AAPN obtained for the global AAPN

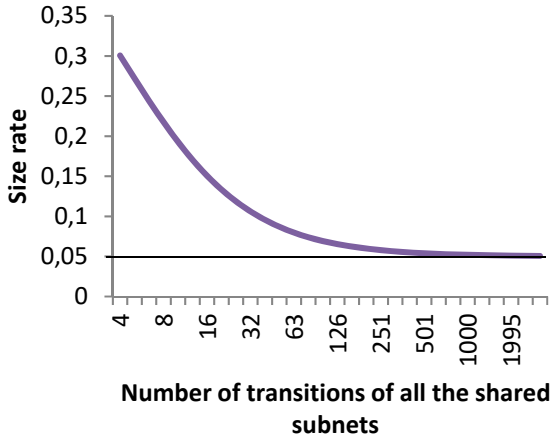


Figure 8. Size rate of global AAPN representing 20 alternative Petri nets.

As an example of the power of reduction of the size of the model achieved by an AAPN, it can be considered that, for example, if the addition of the number of transitions of the four subnets is 200, the size rate of the AAPN that corresponds to the complete model is 0.06. This value means that the size of the AAPN is only 6% of the size of the complete set of 20 alternative Petri nets, despite the fact that the amount of useful information is the same in both models. In other words, 94% of the information contained in the set of 20 alternative PN can be removed to alleviate the computational effort for simulation.

8 Conclusions

The modular construction of a Petri net model of a system has been discussed. This concept has useful application in searching for feasible solutions for a design process of a discrete event system with alternative structural configurations.

The use of an appropriate formalism, such as the AAPN, allows reducing significantly the size of the model that represents all the feasible solutions of the design problem. In the case study presented in this paper, a minimal size rate of 5% can be obtained by the global model, when compared to the original alternative Petri nets. In particular, in the mentioned case study, 95% of the data of the original models is removed in the AAPN. However, the useful information contained in the AAPN is exactly the same as in the original model.

The impact of this research on discrete event systems can be summarized by the fact that in certain cases the design of a system might be developed much faster by a modular construction of alternative solutions and the simulation of compact Petri net models.

References

A.G. Bruzzone and F. Longo. An advanced system for supporting the decision process within large-scale retail stores. *Simulation-Transactions of the Society for Modeling and Simulation International*, 86:742–762, 2010.

R. David and H. Alla. *Discrete, Continuous and Hybrid Petri Nets*. Berlin: Springer, 2005.

E. Jiménez-Macías, and M. Pérez-Parte. Simulation and optimization of logistic and production systems using discrete and continuous Petri nets. *Simulation*, 80(3):143-152, 2004.

J.I. Latorre, E. Jiménez, J. Blanco, and J. C. Sáenz. Optimal Design of an Olive Oil Mill by Means of the Simulation of a Petri Net Model. *International Journal of Food Engineering*, 10(4):573–582, 2014a.

J.I. Latorre, E. Jiménez, and M. Pérez. The optimization problem based on alternatives aggregation Petri nets as models for industrial discrete event systems. *Simulation*, 89(3):346-361, 2013.

J.I. Latorre, E. Jiménez, M. de la Parte, J. Blanco, and E. Martínez. Control of Discrete Event Systems by Means of Discrete Optimization and Disjunctive Colored PNs: Application to Manufacturing Facilities. *Abstract and Applied Analysis*, 2014 Article ID 821707:1-16, 2014b.

J.I. Latorre, E. Jiménez, and M. Pérez. Sequence of decisions on discrete event systems modeled by Petri nets with structural alternative configurations. *Journal of Computational Science*, 5(3):387-394, 2014c.

J.I. Latorre and E. Jiménez. Simulation-based optimization of discrete event systems with alternative structural configurations using distributed computation and the Petri net paradigm. *Simulation*, 89(11):1310-1334, 2013.

J. I. Latorre-Biel, E. Jiménez Macías, J. L. García-Alcaraz, J. C. Sáenz-Díez Muro, M. Pérez de la Parte. Alternatives aggregation petri nets applied to modular models of discrete event systems. In *Proceedings of the European Modelling and Simulation Symposium. EMSS 2015*, pages 465-470. Bergeggi. Italy, 2015.

F. Longo, L. Nicoletti, A. Chiurco, A. O. Solis, M. Massei, and R. Diaz. *Investigating the behavior of a shop order manuf-acturing system by using simulation*. SpringSim EAIA 2013

M. A. Mújica, M.A. Piera, and M. Narciso. Revisiting state space exploration of timed coloured Petri net models to optimize manufacturing system's performance. *Simulation Modelling Practice Theory*, 18:1225–1241, 2010.

M.À. Piera, M. Narciso, A. Guasch, and D. Riera. Optimization of logistic and manufacturing system through simulation: A colored Petri net-based methodology. *Simulation*, 80(3):121-129, 2004.

M. Silva. Introducing Petri nets, F. Di Cesare, editor, *Practice of Petri Nets in Manufacturing*, pages 1-62. Chapman & Hall. 1993.

Transformation of Petri net Models by Matrix Operations

Juan-Ignacio Latorre-Biel [1] Emilio Jiménez-Macías [2] Juan Carlos Sáenz-Díez [2]
Eduardo Martinez-Cámara [2]

[1] Department of Mechanical Energetic and Materials Engineering, Public University of Navarre, Spain,
juanignacio.latorre@unavarra.es
[2] High Technical School of Industrial Engineering, University of La Rioja, Spain,
{emilio.jimenez,julio.blanco,mercedes.perez}@unirioja.es

Abstract

Petri nets constitute a modeling paradigm able to describe discrete event systems characterized by features such as parallelism, precedence, concurrence, and synchronization. Petri nets are applied extensively and successfully for modeling systems belonging to a broad range of fields. In this context, transformation of Petri net models constitutes a process with diverse applications, such as simplifying the model for developing structural analysis or for performance evaluation, as well as comparing different models, describing nets whose structure changes over time, or merging models with exclusive entities. The transformation of the structure of a Petri net can be carried out from different points of view. In this paper, this transformation is developed by means of matrix operations. A list of matrix operations is presented and the preservation of some significant properties of the Petri net is discussed as a practical tool for transforming Petri net models by operations in the incidence matrices.

Keywords: Petri nets, model transformation, matrix-based operations, alternative structural configuration, equivalence class

1 Introduction

Petri nets constitute a paradigm for modeling discrete event systems (DES). They are especially suited for DES characterized by features such as parallelism, precedence, concurrence, and synchronization. A Petri net model of a DES contains different elements for representing both the static structure and the dynamic behavior (Jiménez-Macías E. and Pérez-Parte, 2004; Latorre-Biel and Jimenez-Macias, 2011).

One of the well-known advantages of the Petri nets consists of a double complementary representation. On the one hand, the static structure of a Petri net can be described by means of a directed, weighted, and bipartite graph, whose nodes, called places and transitions, can be classified into two disjoint sets. In this graphical description the dynamics of the DES, and its state, is represented by tokens. On the other hand, the static structure of a Petri net can also be described by means of the input and output incidence matrices. Its state and the dynamic behavior may be represented by a marking or state vector, as well. This matrix-based description provides a quantitative representation of the DES able for structural analysis and performance evaluation (Murata, 1989).

By the application of matrix operations to the incidence matrix of a Petri net, it is possible to change the static structure of the net. This modification leads to a new model of the discrete event system, which may preserve certain properties of the original Petri net, depending on the matrix operation that has been applied. Many properties have been defined for Petri nets, such as equivalence, liveness, reachability, deadlock-freedom, reversibility, soundness, boundedness, controlled siphon property, consistency, conservativeness, persistence, coverability, repetitiveness, or fairness (Murata, 1989; Silva, 1993; Esparza and Nielsen, 1994). Informal descriptions of some of the mentioned properties are (Silva, 1993):

Boundedness. A Petri net is bounded if its set of reachable markings is finite.

Reachability. The reachability problem for Petri nets consists of deciding, given a Petri net (N, M_0) and a marking M of N, if M can be reached from M_0.

Liveness. A Petri net is live if every transition can always occur again.

Deadlock-freedom. A Petri net is deadlock free if every reachable marking enables some transition (Silva, 1993). It is a weaker condition than liveness (Murata, 1989).

Home states. A marking of a Petri net is a home state if it is reachable from every reachable state. The home state problem consists in deciding given a Petri net (N, M_0) and a reachable marking M_0 if M is a home state. The subproblem of deciding if the initial marking of a Petri net is a home state is the problem of deciding if a Petri net is cyclic (Silva, 1993) or reversible (Murata, 1989; Esparza and Nielsen, 1994).

Equivalence. Despite there are different kinds of equivalence, in this paper, the following approach will be followed: two Petri nets are said to be equivalent iff their reachability trees are isomorphous. This property is particularly interesting since it reflects that the Petri

net with isomorphous reachability trees present similar behavior (Medina-Marin et al, 2013; Silva, 1981).

This form of defining equivalence implies the preservation of many properties, since one methodology to prove the verification of certain property by a Petri net consists of analyzing its reachability graph (Murata, 1989). As a consequence, it can be considered as a very useful property, since, for example, performance evaluation depends on the behavior of the system. Furthermore, decision-making support based on the simulation of a Petri net model, is closely related to performance evaluation and, eventually, to the behavior of the system (Silva, 1985).

Transformation of a Petri net constitutes a common process in some applications of this paradigm. For example, certain activities, such as performing structural analysis can be facilitated by the simplification of the Petri net model (Esparza and Nielsen, 1994).

In the case of complex Petri nets or simple Petri nets with complex behavior, calculating the reachable or coverability tree, as well as developing performance evaluation depends strongly on available power of computation. As a consequence, simplifying a model of a system may lead to the use of less computer resources, such as time and memory (Silva and Colom, 1988; Berthelot, 1987).

Matrix based operations may reduce the size of an incidence matrix or may allow comparing two different Petri nets, whose input and output incidence matrices are initially different. Knowing that two Petri nets are equivalent may simplify their analysis, since studying one of them might make unnecessary the analysis of the equivalent Petri net (Silva, 1981).

Moreover, techniques of top-down modeling can be implemented by transforming a low-detailed model into a more refined one (Esparza and Nielsen, 1994).

In order to develop the best transformation for a given purpose it is useful to be aware of the tools available for this task, as well as, which properties are preserved in the Petri net, after the transformation.

Efficient algorithms for transforming Petri net models are performed by modifying their incidence matrices. As a consequence, it constitutes a convenient tool for supporting these processes, taking into account a range of available matrix operations and the properties of the original Petri nets that they preserve (Silva, 1981; Latorre-Biel et al, 2015).

In this paper, some matrix operations applied to the incidence matrix of a Petri net, and the properties preserved in the process, between the original and the resulting Petri nets are discussed.

In the rest of the paper, the following contents are provided. Section 2 discusses the basic concepts of Petri nets. Section 3 discusses some operations and the properties they preserve. Next section addresses the conclusions and future research lines. Last section is related to the bibliographical references.

2 Petri Nets

Some basic concepts of Petri nets are given in this section with the purpose of introducing the reader in the subject and of providing with the notation that will be used in the following section. Moreover, some properties of Petri nets will be discussed. For more information on this topic see (Murata, 1989) and (Silva, 1993).

Following (Murata, 1989), a Petri net is a four-tuple $N = \langle P, T, Pre, Post \rangle$, where, P and T are disjoint, finite, non-empty set of places and transitions respectively, Pre: $P \times T \rightarrow \mathbb{N}$ is the pre-incidence or input function, and $Post$: $T \times P \rightarrow \mathbb{N}$ is the prost-incidence or output function.

The structure of a Petri net can be described by the pre-incidence and post-incidence matrices, W^- and W^+ respectively. In particular, their elements verify that $w_{ij}^+ = Pre(p_i,t_j)$ and $w_{kl}^- = Post(t_l,p_k)$. These values can also be called weight of the arc linking a place and a transition. It can be seen that W^- , $W^+ \in M_{m \times n}$, where $m = |P|$ and $n = |T|$. Every row of the pre-incidence and post-incidence matrices is associated to a different place of the Petri net, while every column is associated to a transition.

It is possible to define a single incidence matrix from the pre-incidence and post-incidence matrices:

$$W = W^+ - W^- \quad (1)$$

There is a characteristic of the Petri net related to the usefulness of the incidence matrix W for describing accurately the static structure of a Petri net. A Petri net is said to be pure if it does not contain any self-loop. Informally, a self-loop is a pair $\{p_i, t_j\}$, where $p_i \in P$ and $t_j \in T$, such that p_i is input place and output place of the same transition t_j. In particular, it can be seen that if $\{p_i, t_j\}$ is a self-loop then $w_{ij}^+ = Pre(p_i,t_j) \neq 0$ and $w_{ij}^- = Post(t_j,p_i) \neq 0$.

Furthermore, given a Petri net $N = \langle P, T, Pre, Post \rangle$, where $|P| = m$ and $|T| = n$, $\forall\, W \in M_{m \times n}\ \exists!\ W^+, W^- \in M_{m \times n}$ such that $W = W^+ - W^-$ iff W is pure. In other words, due to the fact that W^+ and W^- represent the static structure of the Petri net, W is an also valid representation of this structure iff $\nexists\ w_{ij} \in W$, such that $w_{ij} = w_{ij}^- + w_{ij}^+$, where $w_{ij}^+ \neq 0$ and $w_{ij}^- \neq 0$. Notice that otherwise w_{ij} could be decomposed in infinite many positive integers associated to w_{ij}^- and w_{ij}^+.

As we have seen, self-loops reduce the usefulness of the incidence matrix for describing the static structure of a Petri net, since information is lost from W^+, W^- to W. In order to avoid this problem the self-loop can be removed by means of one of the following ways:

a) Introduce a new "dummy" place and transition in one of the arcs of the self-loop (Esparza and Nielsen, 1994). *This solution is general, since it can be applied to any self-loop, but increases the size of the resulting incidence matrix.*

b) If one of the components of the pair that conforms the self-loop have not any additional input or output element, different from the other element of the pair, then, it can be removed. This means that the incidence matrix is reduced in one row or one column. This solution can be applied only to particular cases and it decreases the size of the resulting incidence matrix. Liveness, safeness, and boundedness are preserved (Esparza and Nielsen, 1994).

As a conclusion of the previous considerations, it can be stated that a non-pure Petri net can always be transformed in a Pure Petri net, likely modifying its dimensions. As a consequence, the structure of any Petri net can be described accurately by a single incidence matrix. This conclusion is important for the subsequent considerations on matrix operations applied to incidence matrices of Petri nets.

The previous considerations are related to the static structure of a Petri net. The consideration of its dynamic behavior can be addressed by means of the evolution of the state of the system, which is represented by means of the marking of the Petri net.

The marking of a Petri net can be defined as an application $M: P \rightarrow \mathbb{N}$, assigning a non-negative integer to every place of the net. The state of the Petri net can be represented by a marking or state vector, whose components are the marking associated to every place of the net, $M(p_i)\ \forall\ p_i \in P$.

As a consequence, a marked Petri net or Petri net system, can be defined as the pair $\langle N, M_0 \rangle$, where N is a Petri net and M_0 is its initial marking.

3 Petri Net Analysis

A certain Petri net model can be studied for correctness, following a procedure of qualitative analysis. This approach may allow checking the verification of properties, such as the ones mentioned in the introduction, e.g. liveness, boundedness, or reachability.

There are several techniques for analyzing Petri net systems (Murata, 1989):

- Analysis by enumeration. Requires the construction of the reachability graph for bounded Petri nets or the coverability graph for unbounded Petri nets. With the mentioned graph, it is immediate to prove the verification of properties by the Petri net, such as liveness or reachability. However, the construction of this graph may be computationally costly, even unaffordable, due to the combinatorial state explosion problem.
- Analysis by transformation. This technique is based in the transformation of the Petri net model by the application of operations that preserve the properties expected to prove. This process is aimed at transforming the Petri net into another one, where it is easier to check the verification of certain properties. In fact, this objective may be achieved by arriving to a simpler model or to a Petri net, whose properties are already known.
- Structural analysis. In this case, several techniques of linear programming can be applied to deduce structural properties verified by the Petri net or ad hoc deductions can be performed based on graph-based techniques. As a result, some association between the structure of the Petri net and its dynamic behavior is determined.
- Analysis by simulation. This is the only technique that does not lead to exact results; hence, it cannot allow proving the verification of properties. However, by defining appropriate configurations for the freedom degrees of the Petri net, it is possible to acquire knowledge on the behavior of the system under certain conditions.

On the other hand, the evaluation of the efficiency of the Petri net model is important for certain applications, deducing from this analysis, the efficiency of the modeled discrete event system. This study can be carried out by means of techniques of quantitative analysis or performance evaluation, leading to the calculation of parameters that measure the quality of the system or its behavior, such as yield, costs, or utilization rate, just to give a few examples belonging to the field of manufacturing management.

A transformation of a Petri net model may require preserving certain properties. For example, if the purpose of the transformation is to simplify the model of the system for verifying its correctness, then preserving properties such as liveness or boundedness may be convenient. On the contrary, if the purpose of the transformation is to reduce the computational effort required to evaluate the efficiency of the Petri net, the transformation should lead to a simplified equivalent Petri net in terms of having an isomorphous reachability graph.

In the following section, several matrix operations that can be applied on the incidence matrix of a Petri net model will be presented, as well as some properties they preserve. As a result, a given transformation of a Petri net model could be performed by the application of some of the matrix operations to the incidence matrix of the net, regarding the properties that should be preserved. The implementation of matrix operations in an algorithm is a convenient form of transforming a Petri net in an efficient and automatic process.

4 Matrix Operations

Reference (Silva, 1981) discuss six matrix operations, three for rows and three for columns of the incidence matrix, applied to an incidence matrix; namely:

- **Adding a row (column) of zeros**. The graphical interpretation consists of adding an isolated place (transition) to the Petri net. The dimension of the state or marking vector should be adjusted to the variation in the dimensions of the incidence matrix.
- **Removing a row (column) of zeros**. This operation can be interpreted as removing and isolated place or transition of the Petri net. Analogously to the previous operation,
- **Swapping two rows (columns)**. The interpretation of this operation in the Petri net graph consists of interchanging the names of the places (transitions) associated to the swapped rows (columns) of the incidence matrix. The size of the incidence matrix keeps constant after the application of this operation.

These six matrix operations preserve the structure of the reachability graph (coverability tree for unbounded Petri nets), i.e. both the reachability (coverability) graph of the original and the transformed Petri nets are isomorphous.

As a consequence, it is immediate to prove that the properties of liveness, safeness, boundedness, reachability, reversibility, or equivalence are preserved by the application of one of these matrix operations, as well as of any of their feasible combinations.

A significant effort has already been devoted to the development of reduction rules for Petri nets (Murata, 1989; Esparza and Nielsen, 1994). The application of six of them are discussed from the point of view of matrix operations (Latorre-Biel *et al*, 2015). These six reduction rules lead to some additional elementary matrix operations that can be applied to an incidence matrix for transforming the Petri net model. Any of these six reduction rules preserve properties such as liveness, safeness, and boundedness (Esparza and Nielsen, 1994). As a consequence, the associated matrix operations also preserve these properties. In particular, it can be considered the following matrix operations:

- **Sum of two rows** q and r that verify the properties mentioned below. The resulting row is included in the incidence matrix, while the original rows q and r are removed, as well as the column i, which is also removed. Notice that i is defined in the first of the following properties:
  - *i)* $\exists!\ i \in \mathbb{N}$, where $1 \leq i \leq n = |T|$, such that $w_{qi} \gneqq 0 \wedge w_{ri} \gneqq 0$.
  - *ii*) $\forall\ j \in \mathbb{N}$, where $1 \leq j \leq n = |T|$ and $j \neq i$, it is verified that $w_{qj} = 0 \vee w_{rj} = 0$.
  - *iii*) $\forall\ k \in \mathbb{N}$, where $1 \leq k \leq m = |P|$ and $k \neq q \wedge k \neq r$, then $w_{ki} = 0$.

This operation is based in the reduction rule of fusion of series places or FSP (Esparza and Nielsen, 1994), which is a particular case of the macroplace rule (Latorre-Biel *et al*, 2014 ; Murata, 1989). Property (*i*) justifies the existence of a single common transition between the series places. Property (*ii*) guarantees that the matrix operation does not add any self-loop to the Petri net. In other words, if the original Petri net is pure the matrix operation will lead to a resulting Petri net, which is also pure. Property (*iii*) addresses the fact that the intermediate transition t_i is only linked to the series places. Notice that the column i is associated to the transition t_i.

Analogously it can be stated the following matrix operation:

- **Sum of two columns** q and r that verify the properties mentioned below. The resulting column is included in the incidence matrix, while the original columns q and r are removed, as well as the row i, which is also removed. Notice that i is defined in the first of the following properties:
  - *i)* $\exists!\ i \in \mathbb{N}$, where $1 \leq i \leq m = |P|$, such that $w_{iq} \gneqq 0 \wedge w_{ir} \gneqq 0$.
  - *ii*) $\forall\ j \in \mathbb{N}$, where $1 \leq j \leq m = |P|$ and $j \neq i$, it is verified that $w_{jq} = 0 \vee w_{jr} = 0$.
  - *iii*) $\forall\ k \in \mathbb{N}$, where $1 \leq k \leq n = |T|$ and $k \neq q \wedge k \neq r$, then $w_{ik} = 0$.

The previous matrix operation is based in the reduction rule of fusion of series transitions or FST (Esparza and Nielsen, 1994), which is a particular case of the transition fusion rules (Latorre-Biel and Jiménez-Macías, 2013; Murata, 1989). Property (*i*) justifies the existence of a single common place between the series transitions. Property (*ii*) guarantees that the matrix operation does not add any self-loop to the Petri net. Property (*iii*) addresses the fact that the intermediate place p_i is only linked to the series transitions. Notice that the row i is associated to the place p_i.

- **Removing a row** q if the properties mentioned below are verified:
  - *i)* $\exists\ j, k \in \mathbb{N}$, where $1 \leq j, k \leq n = |T|$, such that $w_{qj} \gneqq 0 \wedge w_{qk} \gneqq 0$.
  - *ii)* $\exists!\ r \in \mathbb{N}$, where $1 \leq r \leq m = |P|$ and $\exists\ j, k \in \mathbb{N}$, where $1 \leq j, k \leq n = |T|$, such that $w_{rj} \gneqq 0 \wedge w_{rk} \gneqq 0$.
  - *iii*) $\forall\ i \in \mathbb{N}$, where $1 \leq i \leq n = |T|$ and $i \neq j \wedge i \neq k$, then $w_{qi} = 0 \wedge w_{ri} = 0$.

This operation is based in the reduction rule of fusion of parallel places or FPP (Esparza and Nielsen, 1994), which is a particular case of the implicit place rule (Latorre-Biel and Jiménez-Macías, 2013b; Latorre-Biel and Jimenez-Macias, 2011). Property (*i*) justifies the existence of an input and an output transition for the place p_q, associated to the *qth* row. Analogously, property (*ii*) guarantees the existence of an input and an output transition for the place p_r, associated to the *rth* row, which are the same as in case of place p_q. Property (*iii*) addresses the fact that there is a single input and

output transitions for both, places p_q and p_r. See figure 1 for a graphical representation.

Analogously it can be stated the following matrix operation:

- **Removing a column** q if the properties mentioned below are verified:
  - *i)* $\exists\, j, k \in \mathbb{N}$, where $1 \leq j, k \leq m = |P|$, such that $w_{jq} \gneqq 0 \wedge w_{kq} \lneqq 0$.
  - *ii)* $\exists!\, r \in \mathbb{N}$, where $1 \leq r \leq n = |T|$ and $\exists\, j, k \in \mathbb{N}$, where $1 \leq j, k \leq m = |P|$, such that $w_{jr} \gneqq 0 \wedge w_{kr} \lneqq 0$.
  - *iii*) $\forall\, i \in \mathbb{N}$, where $1 \leq i \leq m = |P|$ and $i \neq j \wedge i \neq k$, then $w_{iq} = 0 \wedge w_{ir} = 0$.

This operation is based in the reduction rule of fusion of parallel transitions or FPT (Esparza and Nielsen, 1994), which is a particular case of the identical transition rule (Latorre-Biel and Jiménez-Macías, 2013; 2013b). Property (*i*) justifies the existence of an input and an output place for the transition t_q, associated to the *qth* column. Analogously, property (*ii*) guarantees the existence of an input and an output place for the transition t_r, associated to the *rth* column, which are the same as in case of transition t_q. Property (*iii*) addresses the fact that there is a single input and output places for both, transitions t_q and t_r.

It should be mentioned that the opposite matrix operation to the two previously mentioned ones might also been applied: the addition of a row or a column, when there is another one composed of zeros but two elements which present the same absolute value but opposed signs. The added row or column should present the same elements and in the same positions than the one that should be already present in the incidence matrix before the operation. As well as in the previous operations, the properties of liveness, safeness, and boundedness are also preserved.

In particular, it is possible to present the following two matrix operations:

- **Adding a row** q if the properties mentioned below are verified:
  - *i)* $\exists\, j, k \in \mathbb{N}$, where $1 \leq j, k \leq n = |T|$, such that $w_{qj} \gneqq 0 \wedge w_{qk} \lneqq 0$.
  - *ii)* $\exists!\, r \in \mathbb{N}$, where $1 \leq r \leq m = |P|$ and $\exists\, j, k \in \mathbb{N}$, where $1 \leq j, k \leq n = |T|$, such that $w_{rj} \gneqq 0 \wedge w_{rk} \lneqq 0$.
  - *iii*) $\forall\, i \in \mathbb{N}$, where $1 \leq i \leq n = |T|$ and $i \neq j \wedge i \neq k$, then $w_{qi} = 0 \wedge w_{ri} = 0$.

This operation is based in the opposite process to the reduction rule of fusion of parallel places or FPP (Esparza and Nielsen, 1994).

- **Adding a column** q if the properties mentioned below are verified:
  - *i)* $\exists\, j, k \in \mathbb{N}$, where $1 \leq j, k \leq m = |P|$, such that $w_{jq} \gneqq 0 \wedge w_{kq} \lneqq 0$.
  - *ii)* $\exists!\, r \in \mathbb{N}$, where $1 \leq r \leq n = |T|$ and $\exists\, j, k \in \mathbb{N}$, where $1 \leq j, k \leq m = |P|$, such that $w_{jr} \gneqq 0 \wedge w_{kr} \lneqq 0$.
  - *iii*) $\forall\, i \in \mathbb{N}$, where $1 \leq i \leq m = |P|$ and $i \neq j \wedge i \neq k$, then $w_{iq} = 0 \wedge w_{ir} = 0$.

This operation is based in the opposite process to the reduction rule of fusion of parallel transitions or FPT (Esparza and Nielsen, 1994).

The multiplication and the division of elements of an incidence matrix can be useful for transforming a Petri net. Moreover, this operation, as well as the other ones presented in this paper can be applied in sequence to profit from the combination of their effects.

One of the matrix operations that will be analyzed in this section is presented below:

- **Multiplying a row** q **by a positive integer** k. This operation should comply with an additional restriction described in the following. Let us call $M_0(p_k)$ the initial marking of the place p_k, associated to the *qth* column in the original Petri net. Let us call $M_0'(p_k)$ the initial marking of p_k in the Petri net that results from the transformation. It should be verified that:

$$M_0'(p_k) = k \cdot M_0(p_k) \qquad (2)$$

As it has been shown, the application of this matrix operation requires the multiplication of both, a row of the incidence matrix and the initial marking of the place associated to the multiplied row. Graphical interpretation.

It can be proven that the application of this matrix operation preserves the properties of liveness, safeness, and boundedness. This conclusion is immediate when it is realized that the reachability graph (coverability graph if the Petri net is unbounded) is isomorphous in the original and the transformed Petri net.

One of the matrix operations that will be analyzed in this section is presented below:

- **Dividing a row** q **by a positive integer** k. This operation should comply with two additional restrictions described below. Let us call $M_0(p_k)$ the initial marking of the place p_k, associated to the *qth* column in the original Petri net. Let us call $M_0'(p_k)$ the initial marking of p_k in the Petri net that results from the transformation. It should be verified that:
  - *i*) $M_0'(p_k) = M_0(p_k) / k \in \mathbb{Z}$
  - *ii*) $\forall\, i \in \mathbb{N}$, where $1 \leq i \leq n = |T|$ it is verified that $w_{qi} \in \mathbb{Z}$

As it has been shown, the application of this matrix operation requires the division of both, a row of the incidence matrix and the initial marking of the place associated to the multiplied row. The results of these two divisions should be integers, no matter if positive, negative, or zeros.

It can be proven that the application of this matrix operation preserves the properties of liveness, safeness, and boundedness. This conclusion is immediate when it is realized that the reachability graph (coverability graph if the Petri net is unbounded) is isomorphous in the original and the transformed Petri net.

As an example of combined application of matrix operations, to appreciate the advantages of this approach, it can be considered the following matrix operations "dividing a row q by a non-negative integer k" and "removing a column q" it can be considered the following one.

Let us consider the case presented in figure 1. In principle, the reduction rule of fusion of parallel places or FPP is not applicable, since the weight of the arcs of p_q is a and are different from the weight of the arcs of p_r , which is b.

As a consequence, it is possible to divide the qth row of the incidence matrix by b. As it can be seen in figure 1, this row presents only two elements different to zero, whose value is $w_{qj} = b = w_{qk}$. Dividing this row by b will lead to a new qth row, where the zeros are the same, while the other two elements different to zero, in the resulting row present a value of 1.

As a second step, it is possible to apply again the rule of dividing the row r by a, leading to a rth row with zeros, with the exception of $w_{rj} = 1 = w_{rk}$.

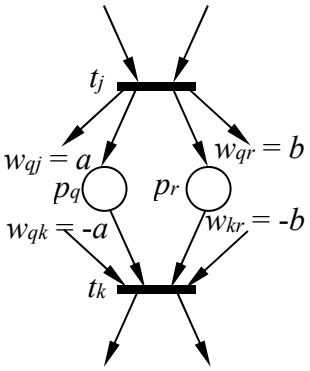


Figure 1. Graphical representation of the initial Petri net in an example of combined application of two matrix operations.

The situation of the Petri net after the application of these two matrix operations can be seen in figure 2.

Now it is possible to apply the matrix operation "removing a row q", since the conditions for the application of this matrix operation are complied.

As a result of this example of application, the size of the Petri net has been decreased in one row, and another row has reduced the values of its elements from $b \in \mathbb{Z}$ to 1.

5 Conclusions

The research line presented in this document has analyzed so far a set of 14 matrix operations for their potential application to the incidence matrix of a Petri net. The application of a sequence of these operations to the incidence matrix of a Petri net allow obtaining another Petri net that preserve certain properties of the initial one.

Special emphasis has been place on the preconditions and the consequences of their application, as well as the properties of the Petri net they preserve. Some of these matrix operations are based on previously developed reduction rules for Petri nets.

These operations can be applied easily in algorithms focused on the transformation of Petri nets for different purposes, such as net analysis, reduction of computer effort devoted to simulating the behavior of the Petri net, the comparison of the incidence matrices of different Petri nets to estimate their similarity, or to merge them into a single compound Petri net, just to give some examples.

As future research lines, it can be considered to increase the number of matrix operations to broaden the range of tools for Petri net transformations, as well as to analyze the preservation of other properties in the already presented matrix operations.

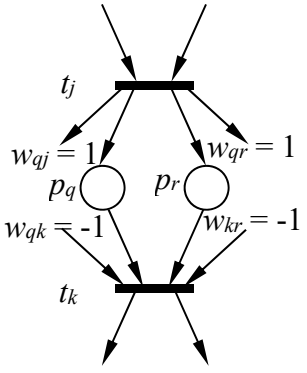


Figure 2. Graphical representation of the final Petri net in an example of combined application of two matrix operations.

References

G. Berthelot. Transformations and decompositions of nets. In *Petri Nets: Central Models and their Properties*. LNCS 254. W. Brauer, W. Reisig, and G. Rozenberg, Eds. Springer-Verlag, pp. 359-376, 1987.

J. Esparza, and M. Nielsen. *Decidability Issues for Petri Nets*. Technical Report, BRICS RS948, BRICS Report Series, Department of Computer Science, University of Aarhus, 1994.

E. Jiménez-Macías, and M. Pérez-Parte. Simulation and optimization of logistic and production systems using discrete and continuous Petri nets. *Simulation*, 80(3):143-152, 2004.

J. I. Latorre-Biel, E. Jiménez Macías, J. L. García-Alcaraz, J. C. Sáenz-Díez Muro, M. Pérez de la Parte. Alternatives aggregation petri nets applied to modular models of discrete event systems". In *Proceedings of the European Modelling and Simulation Symposium. EMSS 2015*, pages 465-470, Bergeggi, Italy, 2015.

J.I. Latorre-Biel, and E. Jimenez-Macias. Matrix-Based Operations and Equivalent Classes in Alternative Petri nets. In *Proceedings of the European Modelling and Simulation Symposium EMSS 2011*, pages 587-592, Rome, 2011.

J.I. Latorre-Biel, and E. Jiménez-Macías. Efficient Methodology For High Level Decision Making On A Manufacturing Facility". In *Proceedings of the 8th EUROSIM Congress on Modelling and Simulation*, pages 345 – 350, Cardiff (United Kingdom), 2013.

J.I. Latorre-Biel, and E. Jiménez-Macías. Simulation-based optimization of discrete event systems with alternative structural configurations using distributed computation and the Petri net paradigm. *Simulation*, 89(11):1310-1334, 2013.

J.I. Latorre-Biel, E. Jiménez-Macías, J. Blanco-Fernández, E. Martínez-Cámara, J.C. Sánez-Díez, and M. Pérez-Parte. Design and operation of a dairy plant by means of a decision support tool based on the Petri nets paradigm. In *Proceedings of the European Modelling and Simulation Symposium EMSS 2014*, pages 588-593, Bordeaux, 2014

J. Medina-Marin, J. C. Seck-Tuoh-Mora, N. Hernandez-Romero, and J. C. Quezada-Quezada. Petri net reduction rules through incidence matrix operations. In *Proceedings of the European Modelling and Simulation Symposium EMSS 2013*, pages 496-503, Athens, 2013.

T. Murata. Petri nets properties analysis and applications. In *Proceedings of the IEEE* 77(4):541–580, 1989. doi: 10.1109/5.24143

M. Silva, and J. M. Colom. On the computation of structural synchronic invariants in P/T nets. In *Advances in Petri nets'88* LNCS 340, G. Rozenberg, Ed. Springer-Verlag, pp. 386-417, 1988.

M. Silva. Introducing Petri nets, In *Practice of Petri Nets in Manufacturing*, F. Di Cesare, (editor), pages 1-62. Ed. Chapman&Hall, 1993.

M. Silva. "*Las redes de Petri en la Automática y la Informática*", translated as Petri nets in automatics and computation. Ed. AC, Madrid, 1985.

M. Silva. Sur le concept de macroplace et son utilisation pour l'analyse des reseaux de Petri". *RAIRO-Systems Analysis and Control*, 15(4):57-67, 1981.

Prediction of Dilute Phase Pneumatic Conveying Characteristics using MP-PIC Method

K. Amila Chandra W.K. Hiromi Ariyaratne Morten C. Melaaen

Faculty of Technology, Natural Sciences and Maritime Sciences – University College of Southeast Norway, Post box 235, N-3603 Kongsberg, Norway,{amila.c.kahawalage, hiromi.ariyaratne, morten.c.melaaen}@usn.no

Abstract

Pneumatic conveying characteristics of a dilute phase flow in a circular horizontal pipe was predicted using MP-PIC method in OpenFOAM code. The geometry, material and operating conditions are similar to some experimental data in published literature. The pipe diameter is 30.5 mm. The solid particles are plastic pellets which are having 1000 kg/m^3 of density and 0.2 mm of particle diameter. The simulations were carried out for 10 m/s of superficial air velocity and for different solids mass loadings 0, 1, 2 and 3. The pressure drop, air velocity profiles and solids distribution were analysed and some of the results were compared with experimental data from the literature. The predicted pressure drops and air velocity profiles show a quite good agreement with the experimental data.

Keywords: MP-PIC, OpenFOAM, pneumatic conveying, simulations, experimental data

1 Introduction

Pneumatic conveying systems are employed to transfer powders, granules and other dry bulk materials through pipes or tubes. The main attractive features of the pneumatic conveying systems are; the flexibility, completely enclosed system and having less moving parts compared to the other mechanical transport systems. One of the principal disadvantages of these systems is the requirement of higher horsepower, because the blower or compressor does the primary work. For better system performance and optimal energy usage, the selected blower or compressor characteristics should be matched with the system characteristics. In that scenario, the air flow rate and the pressure drop through the system are the major key factors when choosing a suitable blower or compressor.

There are two different ways of pneumatic conveying; as dilute phase and dense phase. In the dilute phase conveying, particles are fully suspended in the conveying air. On the other hand in dense phase, the particles are conveyed as fluidized dunes or as discrete plugs of material without much suspension of the material. The remarkable differences in the operational condition for the different modes are; the velocity and the pressure. In dilute phase conveying, relatively a high velocity and a low pressure are employed. Due to high operating velocity in dilute phase, the system requires excessive power. Moreover, operational problems may arise such as particle attrition and erosive wear of the pipelines. The pressure drops for pneumatic conveying systems have widely been measured experimentally by many researchers for different pipe configurations, particle sizes and solids loading ratios (Hyder et al., 2000; Mason and Li, 2000; Tsuji and Morikawa, 1982).

In last few decades, computational fluid dynamics (CFD) is intensively used in modeling, designing and optimizing of pneumatic transport systems. Quite many commercial and open source software programmes are available for that purpose (Bilirgen and Levy, 2001; Chu and Yu, 2008; Hidayat and Rasmuson, 2005; Huber and Sommerfeld, 1998; Laín and Sommerfeld, 2008; Lee et al., 2004; Levy and Mason, 1998; Mason and Levy, 1998).

In general, commercial CFD softwares are user friendly with respect to many aspects such as mesh generation, solution algorithms and visualization. Nevertheless, the modification of source code of those software packages according to user requirement is not very straight forward. Moreover, the costs of commercial licenses are also significant. OpenFOAM is an open source CFD simulation software package and can be used in wide variety of flow simulation applications. It is a finite volume solver. The CFD code can be developed according to the user requirements, as example for a certain specific application. And the code is also for free of charge. Due to the above reasons, OpenFOAM is popular in both academic and industrial sector.

Currently, multiphase particle-in-cell (MP-PIC) method is widely employed in solving gas-solids flow systems. This is also referred as computational particle fluid dynamics (CPFD) in some literature. This is an Euler-Lagrange approach which treats the particles in a discrete manner. Particles are treated as parcels in MP-PIC method and each parcel consists of a definite

number of real particles of the same properties such as size, density, temperature, etc. The method has been quite much used and verified for certain applications such as bubbling and circulating fluidized beds, fluidized bed gasifiers, fluidized beds for carbon capture and gas/liquid/solids fluidized beds (Chen et al., 2013; Karimipour and Pugsley, 2012; Liang et al., 2014; Parker et al., 2013). However, published information about use of this method in predicting pneumatic conveying characteristics is not found.

The solid phase normal stress is used to compute the particle-particle interactions near the close pack limit, but not directly through modeling of particle collisions (Snider, 2001). Besides, the particle collisions are not considered implicitly in MP-PIC method, hence time step size can be increased. Due to that, particle and flow calculation can be computed using same time step size and it reduces the computational time for the simulation. All these benefits make MP-PIC method more suitable for the simulation of the large-scale particulate flow systems. However, in dilute systems the instantaneous and binary contacts are more significant compared to enduring contacts which are modeled through normal stress model. In addition to solid phase normal stress, binary and instantaneous collisions are modeled through new terms developed by Snider and O'Rourke (O'Rourke and Snider, 2012; O'Rourke and Snider, 2010).

In the present study, some of the experimental data found in the literature are reproduced (Tsuji and Morikawa, 1982). Three dimensional simulations are carried out using MP-PIC method in OpenFOAM code. A horizontal circular pipe conveying plastic pellets in dilute phase is simulated. Pressure drop, air velocity profiles and solids distribution are analyzed and some of those results are compared with the experimental data.

2 Model Formulation and Methodology

2.1 Mathematical Model

The mass and momentum equations are solved for the gas phase. For the solid phase, Liouville equation is solved for the distribution function which is a function of particle positions, velocities and sizes (Andrews and O'Rourke, 1996; Snider, 2001). In the equations, refers to the gradient respect to the direction and refers to the gradient respect to the velocity. The mass and momentum equation for the gas phase are shown in (1) and (2), respectively.

$$\frac{\partial(\varepsilon\rho_g)}{\partial t}+\nabla_x(\varepsilon\rho_g u_g)=0 \tag{1}$$

$$\frac{\partial(\varepsilon\rho_g u_g)}{\partial t}+\nabla_x(\varepsilon\rho_g u_g u_g)+\nabla_x p=-F+\varepsilon\rho_g g+\nabla\varepsilon\tau_g \tag{2}$$

The gas phase stress tensor is given by,

$$\tau_g=\mu_{eff}\left(\nabla u_g+\nabla u_g^T\right)-\frac{2}{3}\mu_{eff}\nabla.u_g I \tag{3}$$

Where ε, ρ_g, u_g, p, g, τ_g, μ_{eff}, I are the gas volume fraction (or void fraction), the gas density, the gas velocity vector, the gas pressure, the accerlation due to gravity, the gas stress tensor, effective viscosity and unit tensor, respectively. The turbulent viscosity is solved using the modified k-epsilon equation for multiphase flows (not presented here). The rate of momentum exchange per unit volume from the gas to the particle phase is denoted by (13). The gas pressure and density are correlated by (4).

$$\frac{p}{\rho_g^{\gamma}}=constant \tag{4}$$

The particle phase is described by Liouville equation (5). where $f(x,v,m,t)$ is called the particle distribution function and x,v,m and t represent the particle position, the particle velocity, the particle mass and the time, respectively. More detail about collision term (on the right hand side of (6)) can be found in elsewhere (O'Rourke and Snider, 2012; O'Rourke and Snider, 2010; Snider, 2001).

$$\frac{\partial f}{\partial t}+\nabla_x(fv)+\nabla_v(fA)=\left(\frac{\partial f}{\partial t}\right)_{coll} \tag{5}$$

$$\left(\frac{\partial f}{\partial t}\right)_{coll}=\frac{f_D-f}{\tau_D}+\frac{f_G-f}{\tau_G} \tag{6}$$

The particle velocity is given by,

$$\frac{dx}{dt}=v \tag{7}$$

$A=dv/dt$ is the particle acceleration which is given by

$$A=D(u_g-v)-\frac{1}{\rho_s}\nabla_x p+g-\frac{1}{\theta\rho_s}\nabla_x\tau \tag{8}$$

where D, ρ_s and τ are the drag function, the particle density and the isotropic solids stress, respectively. Drag function is given by (9).

$$D=C_d\frac{3}{8}\frac{\rho_g}{\rho_s}\frac{|u_g-v|}{R} \tag{9}$$

Where C_d is drag coefficient which is modeled from Wen-Yu drag model (Shah et al., 2015) and R is particle radius. Expression for the isotropic solids stress has been taken from (Harris and Crighton, 1994) and shown in (10).

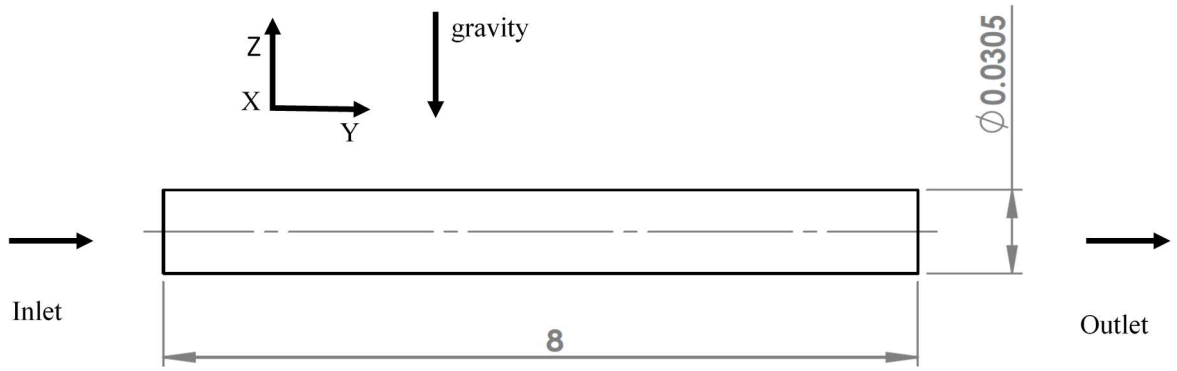


Figure 1. Sketch of the geometry of the computational domain (units are in meters).

$$\tau = P_s \frac{\theta^{\beta}}{\max\left[\theta_{cp} - \theta\ , \omega\left(1-\theta\right)\right]} \tag{10}$$

In (10), P_s is a constant with units of pressure and θ_{cp} is the particle-phase volume fraction at close pack, respectively. β is a constant (2≤β≤5) and ω is a small number in the order of 10-7.The particle volume fraction is related to the distribution function by,

$$\theta = \iint f \frac{m}{\rho_s} dmdv \tag{11}$$

Then ε and θ are related by,

$$\theta + \varepsilon = 1 \tag{12}$$

To complete the equation, we need an expression for the interphase momentum transfer function F and it is defined as (13).

$$F = \iint fm\left[D\left(u_G - v\right) - \frac{1}{\rho_s}\nabla p \right] dmdv \tag{13}$$

2.2 Experiment and Simulation Procedure

Tsuji and Morikawa (1982) have conducted experiments for gas-solids two phase flow in a horizontal pipe. The pipe diameter is 30.5 mm. The pressure drops and also particle and air velocities have been measured using laser-Doppler velocimeter (LDV). Plastic pellets which are having particle density of 1000 kg/m^3 have been used as the solid material.

They have conducted experiments for two different mean particle sizes; 0.2 mm and 3.6 mm by varying superficial air velocity and solids mass loading. The velocities range from 6 to 20 m/s and the solids mass loadings range from 0 to 6. However in the present study, the simulations are conducted for mean particle size of 0.2 mm. The used air density and viscosity are 1.225 kg/m^3 and 1.46073×10^{-5} Pa s, respectively. The simulations are carried out for four different mass loadings; 0, 1, 2 and 3. The superficial air velocity for each case is 10 m/s. The Reynolds number of the flow is around 21000. Description of each simulation case is shown in Table 1.

Table 1. Simulation case description.

Case	*Solids to air mass flow ratio*	*Air mass flow rate (kg/s)*	*Solids mass flow rate (kg/s)*	*Particle volume fraction at inlet (%)*
Case_0	0	0.009	0.000	0.00
Case_1	1	0.009	0.009	0.12
Case_2	2	0.009	0.018	0.24
Case_3	3	0.009	0.027	0.32

2.3 Geometry and Meshing

The computational domain is an 8 m long horizontal pipe having 30.5 mm of diameter. This is shown in Figure 1. Three-dimensional geometry was generated and meshed using SALOME 7.5.1. To obtain better accuracy and convergence, hexahedral type elements were selected. The grid is uniform and consists of 126000 elements (Figure 2).

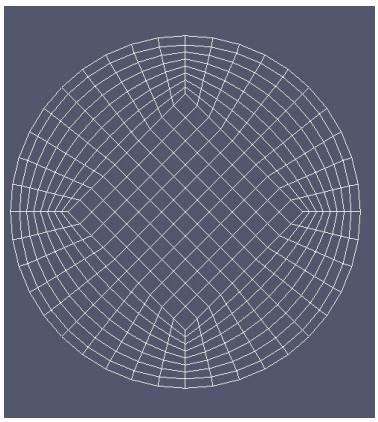
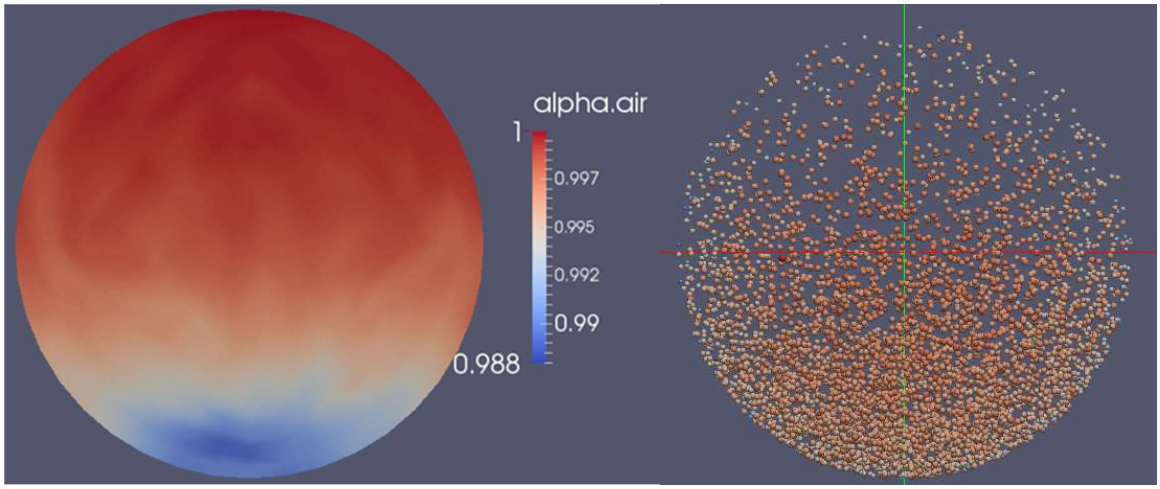


Figure 2. Mesh of pipe cross section.



(a) (b)

Figure 3. (a) Air volume fraction and (b) Particle distribution at outlet after 1.8 s for the Case_2.

2.4 Boundary Conditions and Solution Method

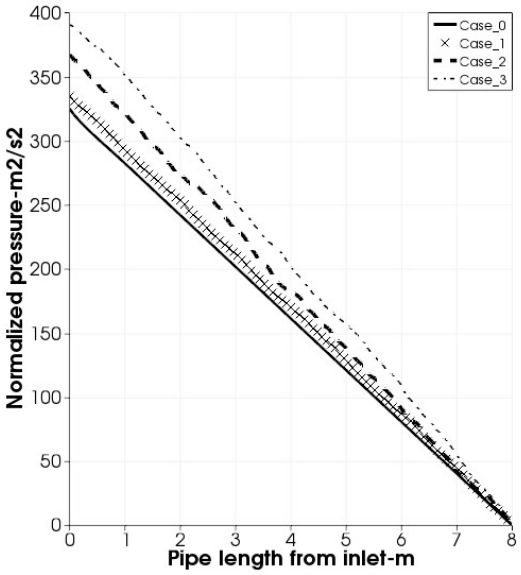


Figure 4. Normalized pressure along the pipe centre line at 1.8 s.

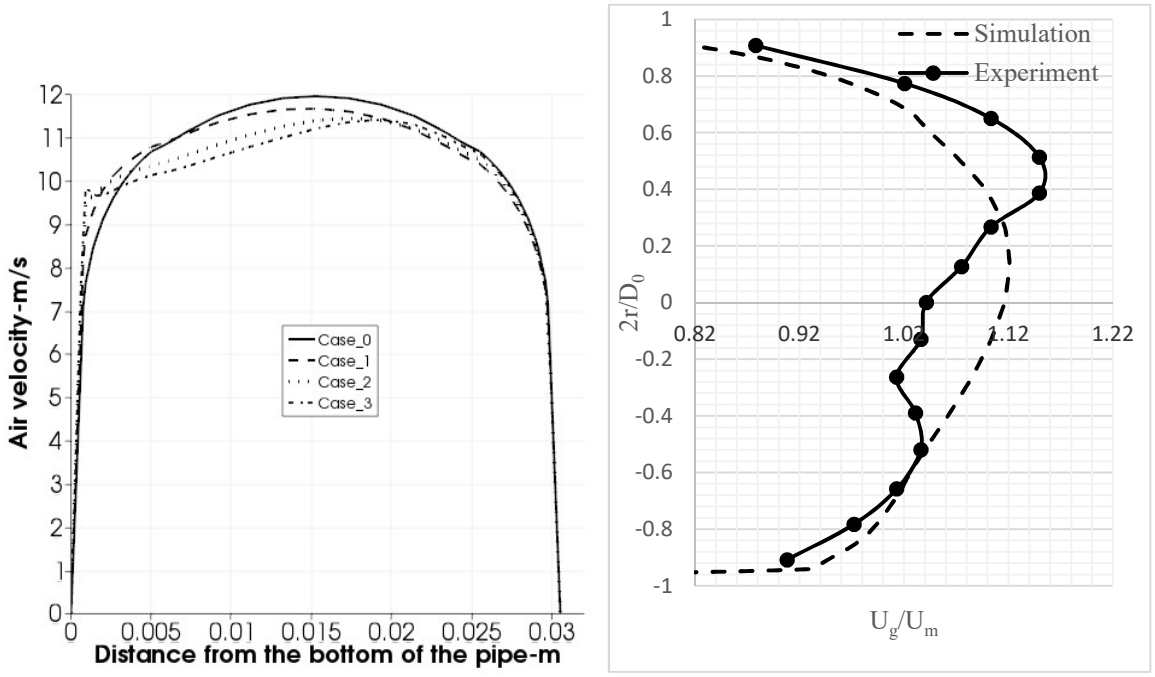
As shown in Figure 1, the system consists of three boundaries as the inlet, the outlet and the wall. Pressure at the outlet was defined as zero gauge pressure and superficial air velocity at the inlet was defined as 10 m/s. Particle-particle interactions nearby close pack limit are modeled using particle normal stress model.

The isotropic solids stress was defined as (10) and P_s , θ_{cp} , β and ω are specified as 1, 0.6, 3 and 10^{-8}, respectively. Particle to wall interaction was modeled with restitution coefficients and its value is 0.95. Collisional return-to-isotropy of particle velocity fields which are important for dilute systems (O'Rourke and Snider, 2012) are modeled from time scale model.
In that model, particle-phase volume fraction at close pack and particle-particle restitution coefficient are specified as 0.6 and 0.95, respectively. MPPICFoam was used as the solver in OpenFOAM. Simulation was run in transient mode at time step size of 0.0001 s until it comes to quasi-steady state which was confirmed by monitoring the pressure at certain points.

3 Results and Discussion

The pressure drop in a pneumatic conveying system is a very crucial property because it will affect to the performance of the blower or the compressor. Pressure drops along the horizontal pipe center line at 1.8 s (normalized by the air density) for four cases are shown in Figure 4. Case_0 which corresponds to only air flow shows quite stable pressure fall along the pipe length; however the pressure drop profiles start to fluctuate with the solids loadings. Moreover, the total pressure drop increases with the increase of solids loadings, which is physically reasonable for any dilute system. When the solids loading increases for a certain size of particles, the particle number density in the system increases accordingly. Higher number of particles in the system causes high frequent particle-particle and particle-wall collisions. This enhances the particle energy dissipation resulting in high drag force and also increased pressure drop.

Table 2 shows the comparison of experimental data (Tsuji and Morikawa, 1982) with predicted quasi-steady time-averaged pressure drops in fully developed region along the pipe centre line for different cases. The simulation results show quite good agreement with the experimental data. However, it seems that the error is increased with an increase of solids loadings. It should be noted that the pressure drops are simulated with



(a) (b)

Figure 5. (a) Air velocity profiles for each case at outlet at 1.8 s and (b) Comparison of experimental and predicted air velocity profiles at outlet at 1.8 s for the Case_2.

certain set of particle-particle and particle-wall collision coefficients in the present study and some of the literature emphasizes the high sensitivity of some of these coefficients on pressure drop results (Patro and Dash, 2014a, 2014b). However, the investigation of this effect is out of the present study.

Table 2. Quasi-steady time-averaged pressure drop in fully developed region along the pipe center line.

Case	Pressure drop (Pa/m)		Error (%)
	Experimental	Simulation	
Case_0	52.63	49.4	-6.1
Case_1	59.6	53.0	-11.2
Case_2	66.2	58.2	-12.1
Case_3	73.5	62.4	-15.1

Figure 3(a) and Figure 3(b) show the air volume fraction and solids distribution, respectively, at outlet after 1.8 s for the case 2. The air volume fraction is lower nearby bottom of the pipe (Figure 3 (a)) because more solids are dominated in that region (Figure 3(b)). Because of the gravity effect, the particles gradually pass off-axis and decline. Therefore, the particles tend to move to bottom side of the pipe. However, due to particle-wall collisions the particles bounce back to the core region of the pipe and are more scattered due to further particle-particle and particle-wall collisions (Figure 3(b)). Still the volume fraction of solids in bottom part of the pipe is higher compared to the upper part.

Figure 5 (a) shows the air velocity profiles at outlet at 1.8 s for each simulated case. Case_0 which corresponds to only air flow shows a symmetric profile. Also, it seems that the model predicts the turbulent air velocity profile quite accurately. With increase of solids loadings, the profiles pronounce the asymmetry i.e. the air velocity is getting lower in the bottom part of the pipe compared to the upper part. This can be due to more restriction for the air flow caused by high amount of solids at the bottom. However, the results in the region nearby the bottom (around 0.001m in Figure 5(a)), in where the air velocities are higher for the solids loading cases (case_1, Case_2 and Case_3) than only air case (Case_0), is not as expected.

Figure 5(b) shows the comparison of experimental and simulated velocity profiles for the case 2. The agreement between two profiles nearby bottom part of the pipe seems reasonable. However, the local minimum and maximum of the experimental profile have not been captured by the model. This can be partly due to mono-dispersed particles used in the simulations in contrast to poly-dispersed particles used in the experiments. Moreover, calibration of particle-wall collision parameters might be necessary for more accurate results.

4 Conclusions

The flow characteristics of a pneumatic conveying system are predicted using MP-PIC method in OpenFOAM. The flow is dilute and plastic pellets which are having 0.2 mm mean particle size and 1000 kg/m3

are conveyed in 30.5 mm diameter pipe. The superficial air velocity was 10 m/s and predictions are performed for 4 different solids loadings as 0, 1, 2 and 3. Some of the predicted results were compared with experimental data. Predicted pressure drop results have reasonable agreement with experimental data for different loadings. The error ranges from 6-15%. The solids distribution also seems physical, however no real world data is available for the comparison. Asymmetry of the gas phase velocity of the profiles due to presence of the particles have been quite well predicted by the model, however the local minimum and maximum of the experimental profile has not been captured by the present model. The discrepancies may be partly due to mono-dispersed particles used in the model in contrast to poly-dispersed particles used in the experiments. In general, it can be concluded that the used MP-PIC model gives quite reasonable predictions for dilute phase pneumatic conveying systems.

Acknowledgement

The authors would like to acknowledge the financial support provided by the Research Council of Norway under PETROMAKS II program and Det Norske oljeselskape ASA.

Nomenclature

A	particle accerlation (m/s^2)
C_D	drag coefficient
D	darg funnction (s-1)
D_0	pipe diameter (m)
F	rate of momentum exchange per unit volume from the gas to the particle phase (N/m3)
f	particle distribution function (PDF)
f_D	PDF obtained by collapsing the velocity dependence f of to a delta function centered about the local mass-averaged particle velocity
f_G	equilibrium distribution
I	unit tensor
m	particle mass (kg)
p	static pressure (Pa)
P_s	pressure constant (Pa)
R	particle radius (m)
r	vertical distance from pipe horizontal axis (m)
t	time (s)
u_g	gas velocity vector (m/s)
Ug	axial gas velocity (m/s)
Um	superficial air velocity at inlet (m/s)
v	particle velocity vector (m/s)
x	particle position (m)
β	constant
γ	constant
ε	gas volume fraction
θ	solids volume fraction
θ_{cp}	particle phase volume fraction at close pack
μ_{eff}	effective viscosity (Pa s)
ρ_g	gas density (kg/m3)
ρ_s	particle density (kg/m3)
τ	isotropic solids stress (Pa)
τ_D	collision damping time (s)
τ_g	gas stress tensor (Pa)
τ_G	relaxation time (s)
ω	constant

References

M. J. Andrews and P. J. O'Rourke. The multiphase particle-in-cell (MP-PIC) method for dense particulate flows. *International Journal of Multiphase Flow, 22*(2): 379-402, 1996. doi:10.1016/0301-9322(95)00072-0.

H. Bilirgen and E. K. Levy. Mixing and dispersion of particle ropes in lean phase pneumatic conveying. *Powder Technology, 119*(2–3): 134-152, 2001. doi:10.1016/S0032-5910(00)00413-7.

C. Chen, J. Werther, S. Heinrich, H.-Y. Qi, and E.-U. Hartge. CPFD simulation of circulating fluidized bed risers. *Powder Technology, 235*: 238-247, 2013. doi:10.1016/j.powtec.2012.10.014.

K. W. Chu and A. B. Yu. Numerical simulation of complex particle–fluid flows. *Powder Technology, 179*(3): 104-114, 2008. doi:10.1016/j.powtec.2007.06.017.

S. E. Harris and D. G. Crighton. Solitons, solitary waves, and voidage disturbances in gas-fluidized beds. *Journal of Fluid Mechanics, 266*: 243-276, 1994. doi:10.1017/S0022112094000996.

M. Hidayat and A. Rasmuson. Some aspects on gas–solid flow in a U-bend: Numerical investigation. *Powder Technology, 153*(1): 1-13, 2005. doi:10.1016/j.powtec.2005.01.016.

N. Huber and M. Sommerfeld. Modelling and numerical calculation of dilute-phase pneumatic conveying in pipe systems. *Powder Technology, 99*(1): 90-101, 1998. doi:10.1016/S0032-5910(98)00065-5.

L. M. Hyder, M. S. A. Bradley, A. R. Reed, and K. Hettiaratchi. An investigation into the effect of particle size on straight-pipe pressure gradients in lean-phase conveying. *Powder Technology, 112*(3): 235-243, 2000. doi:10.1016/S0032-5910(00)00297-7.

S. Karimipour and T. Pugsley. Application of the particle in cell approach for the simulation of bubbling fluidized beds of Geldart A particles. *Powder Technology, 220*: 63-69, 2012. doi:10.1016/j.powtec.2011.09.026.

S. Laín and M. Sommerfeld. Euler/Lagrange computations of pneumatic conveying in a horizontal channel with different wall roughness. *Powder Technology, 184*(1): 76-88, 2008. doi:10.1016/j.powtec.2007.08.013.

L. Y. Lee, T. Yong Quek, R. Deng, M. B. Ray, and C.-H. Wang. Pneumatic transport of granular materials through a

bend. *Chemical Engineering Science, 59*(21): 4637-4651, 2004. doi:10.1016/j.ces.2004.07.007.

A. Levy and D. J. Mason. The effect of a bend on the particle cross-section concentration and segregation in pneumatic conveying systems. *Powder Technology, 98*(2): 95-103, 1998. doi:10.1016/S0032-5910(97)03385-8.

Y. Liang, Y. Zhang, T. Li, and C. Lu. A critical validation study on CPFD model in simulating gas–solid bubbling fluidized beds. *Powder Technology, 263*: 121-134, 2014. doi:10.1016/j.powtec.2014.05.003.

D. J. Mason and A. Levy. A comparison of one-dimensional and three-dimensional models for the simulation of gas–solids transport systems. *Applied Mathematical Modelling, 22*(7): 517-532, 1998. doi:10.1016/S0307-904X(98)00002-X.

D. J. Mason and J. Li. A novel experimental technique for the investigation of gas–solids flow in pipes. *Powder Technology, 112*(3): 203-212, 2000. doi:10.1016/S0032-5910(00)00294-1.

P. J. O'Rourke and D. M. Snider. Inclusion of collisional return-to-isotropy in the MP-PIC method. *Chemical Engineering Science, 80*: 39-54, 2012. doi:10.1016/j.ces.2012.05.047.

P. J. O'Rourke and D. M. Snider. An improved collision damping time for MP-PIC calculations of dense particle flows with applications to polydisperse sedimenting beds and colliding particle jets. *Chemical Engineering Science, 65*(22): 6014-6028, 2010. doi:10.1016/j.ces.2010.08.032

J. Parker, K. LaMarche, W. Chen, K. Williams, H. Stamato, and S. Thibault. CFD simulations for prediction of scaling effects in pharmaceutical fluidized bed processors at three scales. *Powder Technology, 235*: 115-120, 2013. doi:10.1016/j.powtec.2012.09.021.

P. Patro and S. K. Dash. Numerical Simulation for Hydrodynamic Analysis and Pressure Drop Prediction in Horizontal Gas-Solid Flows. *Particulate Science and Technology, 32*(1): 94-103, 2014a. doi:10.1080/02726351.2013.829543.

P. Patro and S. K. Dash. Prediction of acceleration length in turbulent gas–solid flows. *Advanced Powder Technology, 25*(5): 1643-1652, 2014b. doi:10.1016/j.apt.2014.05.019

S. Shah, K. Myöhänen, S. Kallio, and T. Hyppänen. CFD simulations of gas–solid flow in an industrial-scale circulating fluidized bed furnace using subgrid-scale drag models. *Particuology, 18*: 66-75, 2015. doi:10.1016/j.partic.2014.05.008.

D. M. Snider. An Incompressible Three-Dimensional Multiphase Particle-in-Cell Model for Dense Particle Flows. *Journal of Computational Physics, 170*(2): 523-549, 2001. doi:10.1006/jcph.2001.6747.

Y. Tsuji and Y. Morikawa. LDV measurements of an air–-solid two-phase flow in a horizontal pipe. *Journal of Fluid Mechanics, 120*: 385-409, 1982. doi:10.1017/S002211208200281X.

Simulation of Flame Acceleration and DDT

Knut Vaagsaether
Department of Process, Energy and Environmental Technology, University College of Southeast Norway, Porsgrunn, Norway, knutv@usn.no

Abstract

This paper presents a combustion model and a simulation method for modeling flame acceleration (FA) and deflagration to detonation transition (DDT) in a premixed gas. The method is intended to produce the most important effects in FA and DDT without resolving the flame front on the computational mesh. The simulations presented here are of stoichiometric hydrogen-air mixtures in a channel with repeated obstacles. The channel is 2 m long and 110 mm wide, with a height of either 20 mm or 40 mm. The obstacles gives a blockage ratio of 0.5. These values are the same as for experiments by other researchers and is used for comparison. The combustion model combines a turbulent burning velocity model and a two-step Arrhenius kinetic rate. The simulations show similar flame speeds and pressures as seen in experiments, and the process of DDT is shown to be caused by shock focusing and shock flame interactions. The simulations show that the quasi detonation regime is a series of transition to detonation events followed by failure of the detonation. Results from both 2D and 3D simulations are presented, since the 2D simulations show how the method can reproduce important effects.

Keywords: CFD, flame acceleration, DDT, detonation, hydrogen

1 Introduction

Simulations of strong flame acceleration and deflagration to detonation transition (DDT) in gaseous mixtures are important for understanding the propagation of gas explosions. Simulation tools are also important for risk assessment in industries where gas explosions might occur. Gamezo et al. (2007) presented simulations of flame acceleration and DDT in obstructed channels (Gamezo et al., 2007), in which a single step Arrhenius reaction rate describes the chemistry. The computational mesh resolution in their simulations was approximately 100th of a flame thickness. At a larger scale, such resolutions might be impossible to accomplish. Strong flame acceleration in a complex geometry is usually a product of classical fluid mechanical instabilities such as turbulence, Kelvin-Helmholtz, Rayleigh-Taylor, and Richtmyer-Meshkov. Other important effects are the flame-geometry and pressure wave-geometry interactions, in which the flame surface area increases or the reactants are compressed. With sufficient compression, the hot pockets in the reactants might ignite and possibly lead to DDT. Lee and Moen (1980) and Lee et al. (1985) described different propagation regimes in obstructed channels, including: i) choked flow, ii) quasi-detonation, and iii) Chapman-Jouguet (CJ) detonation. The choked flow regime is a deflagration in which the expansion of gas over an obstacle produces high flame speeds. In the quasi-detonation regime, the flame undergoes a transition to detonation followed by a failure of detonation due to diffraction. This DDT and failure process is repeated to produce higher average flame speeds than in the choking regime but lower than the CJ detonation speed. Thomas (2012), Shepherd (2009), Ciccarelli and Dorofeev (2008), and Shepherd and Lee (1992) have written excellent reviews on flame acceleration and DDT. The objective of this work is to develop a simulation method to predict the strong flame acceleration and DDT that can occur in gas explosions in channels with repeated obstacles. The focus of this paper is the development of a combustion model that can reproduce the phenomena seen in flame acceleration and DDT on an under-resolved computational mesh. More details on the method presented in this paper are described in Vaagsaether (2010).

2 Combustion Model

The combustion model combines a turbulent burning velocity model with a chemical kinetic rate model. A two-step reaction model is used for the chemical kinetic rate in all cases. Two reaction variables describe the total reaction, one for the initiation and chain branching reactions, and one for the termination rates. It is assumed that the initiation and branching reactions are isothermal. Three species are conserved as two reaction progress variables. These species are reactants, intermediates (radicals), and products. Here, the two progress variables are called α and β and the conservation equations are provided in (1) and (2).

$$\frac{\partial \rho \alpha}{\partial t} + \nabla \cdot (\rho \vec{U} \alpha) = \dot{\Omega} \quad (1)$$

$$\frac{\partial \rho \beta}{\partial t} + \nabla \cdot (\rho \vec{U} \beta) = \dot{\omega} \quad (2)$$

Since the reaction described by (1) is assumed isothermal and the products of that reaction are assumed to be radicals, the reaction term for (1) is given by the induction time as seen in (3).

$$\dot{\Omega} = \frac{\rho}{\tau_{ind}} \quad (3)$$

Equation (4) shows a typical form for the induction time model, where $[F]_0$ and $[O_2]_0$ are the non-reacted fuel and oxygen concentrations, A_{ind} is a pre-exponential factor, and $T_{a,ind}$ is an activation temperature.

$$\tau_{ind} = A_{ind}[F]_0^x [O_2]_0^y exp(T_{a,ind}/T) \quad (4)$$

$$\dot{\omega} = \max[(\rho_u S_T |\nabla \beta|), (\dot{\omega}_k)] \quad (5)$$

The reaction term in (2) accounts for both turbulent reactions and chemical kinetics. The rate is the maximum reaction rate of a turbulent burning velocity model and an Arrhenius-type kinetic rate. This term is seen in (5) where the first part of the rate is a typical flame density model and the second part is a Arrhenius kinetic rate.

$$\dot{\omega}_k = \begin{cases} A\rho^n T^m \beta^o exp\left(-\frac{T_a}{T}\right) & if\ \alpha \geq 1 \\ 0 & if\ \alpha < 1 \end{cases} \quad (6)$$

Since the turbulent burning velocity model contains all reactions in the flame front, unlike the kinetic term it is not dependent on α. A typical form of the kinetic rate of β is shown in (6), where A is a pre-exponential factor and T_a is an activation temperature. In this model the α-reaction must finish before the Arrhenius rate in the β-reaction can start, as seen in (6). In the unreacted unheated gas, α is 0 and increases with increasing temperature. In gas mixtures with a low temperature in the unreacted gas, the turbulent rate is dominant and the model behaves like a turbulent burning rate model. When the temperature in the reactants increases and the induction time becomes sufficiently small, the kinetic rate in the β equation will start to influence the total rate. This property of the model can capture the effect of ignition by shock compression of the reactants.

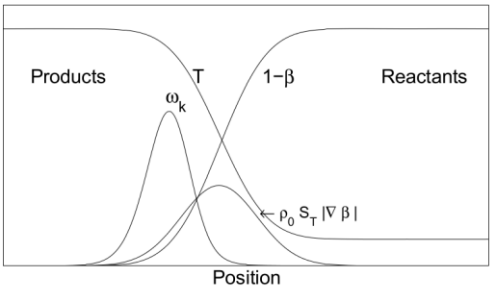


Figure 1: Schematic representation of the reaction rates with β and temperature curves. The Turbulent rate highest at the highest gradient of the reaction rate variable. The Arrhenius rate is highest towards the product side.

Figure 1 shows a schematic representation of the combustion model with the two reaction rates for turbulent combustion and the Arrhenius rate. When the temperature in the reactants increases, the peak in the Arrhenius rate moves toward the reactant side.

3 Geometry Simulation set-up

Figure 2 shows a schematic of the experimental setup from Teodorczyk (2007), which is also the present simulation domain. The channel is 2 m long, 110 mm wide, and closed in all directions. The blockage ratio for all experiments is 0.5. The ignition of the stoichiometric hydrogen-air mixture at atmospheric pressure and 293 K occurs at the center of one end wall. Results from the simulations with channel heights of 20 mm and 40 mm are presented here. These simulations are for a 2D geometry and show that the model handles the most important effects of flame acceleration and transition to detonation in this type of geometry. Two different mesh sizes, 1 mm and 0.5 mm, are tested, but most of the results presented here use the 1 mm mesh. Some results from the 3D simulations are shown and compared with the experimental results. The 3D simulations use a constant 1 mm mesh. Since the real geometry is 3D, the shock focusing and reflections behave differently than they would in 2D.

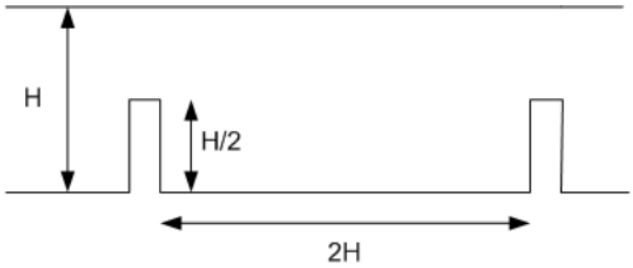


Figure 2: Experimental set-up of Teodorczyk with channel height and distance between obstacles.

3.1 3.1 Simulation set-up

A second-order centered TVD method (FLIC) solves the transport equations for mass, momentum and energy and the gas is modeled as an ideal gas. This method is described in Toro (1999). A turbulence model is used since the flow is turbulent and the turbulent length scales are smaller than the mesh size. Equation (7) shows the one-equation model for the turbulent kinetic energy used in the simulations, where [R] is the Reynolds stress, [S] is the mean strain rates, C_e is a constant set to 0.92, and the turbulent length scale l is set to be the mesh size. The notation ":" is the Frobenius inner product of two matrices.

$$\rho\frac{Dk}{Dt} = \nabla\cdot(\mu_t\nabla k) - [R]:[S] - C_e\rho\frac{k^{1.5}}{l} \tag{7}$$

The turbulent burning velocity is calculated from a model presented in (Flohr and Pitsch, 2000) and is shown in (8).

$$S_T = S_L\left(1 + A\frac{(Re\cdot Pr)^{0.5}}{Da^{0.25}}\right) \tag{8}$$

A is a model constant and is set to 0.52, Re is the turbulent eddy Reynolds number, Pr is the Prandtl-number, Da is the turbulent eddy Dahmkohler number and the velocity fluctuation is calculated as u'=(2/3 k)$^{0.5}$.
The induction time model for hydrogen-air is presented in Sichel et al. (2002) , where A_{ind}=1.1085·10^{10} Pa·s/K for a stoichiometric mixture.

$$\tau_{ind} = A_{ind}\frac{T}{p}\exp(B) \tag{9}$$

$$B = -35.17 + \frac{8530.6}{T} + 7.22\cdot10^{-11}\left(\frac{p}{p_0}\right)^2 exp\left(\frac{21205}{T}\right) \tag{10}$$

The rate of the exothermic reaction was presented in (Korobeinikov, 1972) where A_β=1.04·10^{-5} and E_a=2000 K.

$$\frac{\partial\beta}{\partial t} = A_\beta p^2 exp\left(-\frac{E_a}{T}\right)\left(\beta^2 - (1-\beta)^2 exp\left(-\frac{Q}{RT}\right)\right) \tag{11}$$

The laminar burning velocity is taken from a model presented in Iijima and Takeno (1984), where p_{0s}=101325 Pa and T_{0s}=291 K.

$$S_L = S_{L,0}\left(1 + X\cdot log_{10}\left(\frac{p}{p_{0s}}\right)\right)\left(\frac{T}{T_{0s}}\right)^Y \tag{12}$$

$$S_{L,0} = 2.98 - (\varphi - 1.7)^2 + 0.32(\varphi - 1.7)^3 \tag{13}$$

$$X = 0.43 + 0.003(\varphi - 1) \tag{14}$$

$$Y = 1.54 + 0.026(\varphi - 1) \tag{15}$$

The total energy per volume is:

$$E_{tot} = \frac{p}{\gamma - 1} + \frac{1}{2}\rho\vec{U}\cdot\vec{U} + \rho(1-\beta)Q \tag{16}$$

The properties Q = 3.2 MJ/kg, γ_u=1.402, γ_b=1.242 are used for the stoichiometric mixture of hydrogen-air, where the subscripts u and b indicate the unburned and burned states.

4 Results and discussion, 2D simulation

To demonstrate how the simulation method handles shock ignition and DDT. Two-dimensional simulations are presented and discussed in this section. These results are presented as contour plots of the density gradients as well as plots of the flame speeds along the channel length just below the top wall. Fig. 3 shows the flame speed along the length of the 40 mm high channel. From ignition, the flame speed increases as the flame passes the obstacles. The expansion of the gas across the obstacles produces high flame speed and increases the burning rates, which in turn produce a shock wave traveling in front of the flame. Figure 4 shows the process of DDT in the 40 mm channel. When the shock passes an obstacle, a diffracted shock front reflects at the bottom wall and creates a Mach stem. Both the leading shock and the Mach stem reflect at the obstacles and are focused in the corner between the bottom wall and the obstacle. They ignite the gas behind the focused shock to send a strong shock wave into the products. This shock wave diffracts over the obstacles and reflects at the top wall. The reflected and diffracted shock interacts with the flame from the product side and accelerates the flame, and it may even heat the reactants in the front of the flame to cause DDT.

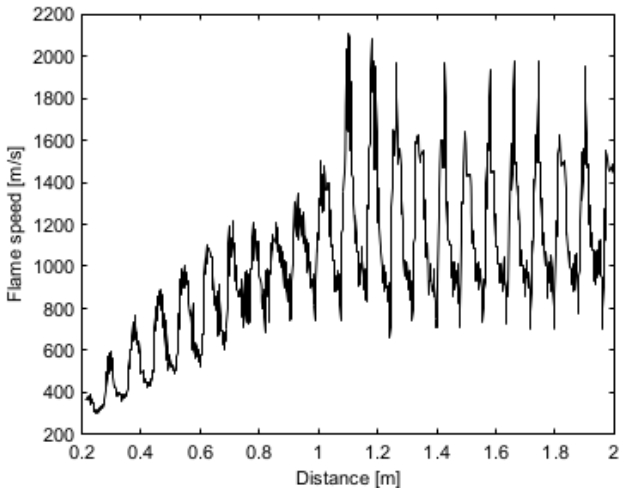


Figure 3: Simulated flame speed as a function of time for the 40 mm channel with repeated obstacles with 1 mm mesh. Stoichiometric hydrogen-air at 293 K and 1 atm.

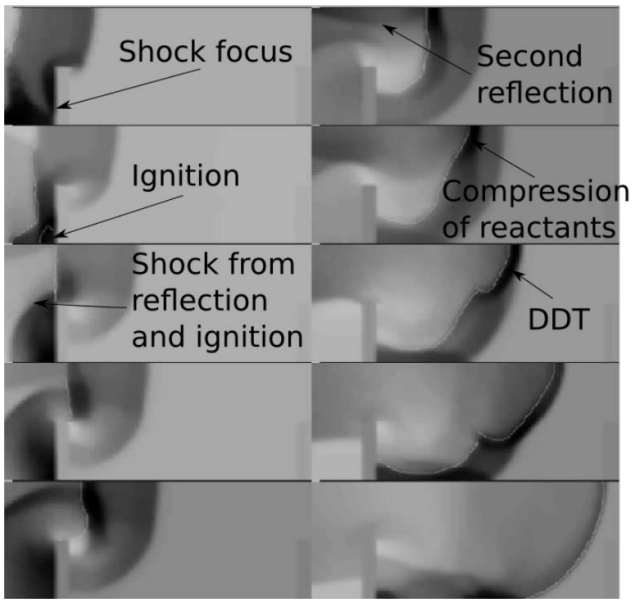


Figure 4: 2D simulation of 40 mm channel height showing density gradients of shock-flame-obstacle interactions where transition to detonation occurs.

When the detonation in the 40 mm channel propagates past an obstacle, the shock diffracts and the detonation fails, as can be seen in Figure 5. Due to the diffraction over an obstacle, the flame does not propagate as a CJ detonation. The average flame speed after this point is about 1400 m/s and can be interpreted as the quasi-detonation regime. In the experiments, there is significant scatter in the locations of the first DDT. The average flame speed in the experiments after this first transition is about 1250 m/s, but varies as much as 200 m/s. A similar process of DDT and failure is seen in high-speed photographs in experiments in Teodorczyk et al., (1988).

The simulations of the 20 mm channel show similar flame acceleration to what occurred in the 40 mm channel. Figure 6 shows the simulated flame speed along the channel length. After about 0.4 m distance from ignition, the flame reaches an average speed of over 1000 m/s, fluctuating between 1200 m/s and 800 m/s, and it is described as the choking regime for this case. In the experimental results there is likely a transition to detonation around 0.7 m, which is not seen in the simulation. But after about 1.0 m the flame speed is on average constant around 1000 m/s in both the experiments and in the simulations. The coarse mesh is unable to resolve the smaller scales of the different instabilities important in flame acceleration, and these instabilities may form small hotspots that can cause DDT. The results of a grid sensitivity test are shown in Figure 7. The 40 mm channel is probably the most interesting case since it includes flame acceleration, DDT, and failure. The flame speed along the 40 mm channel is roughly the same for both mesh sizes.

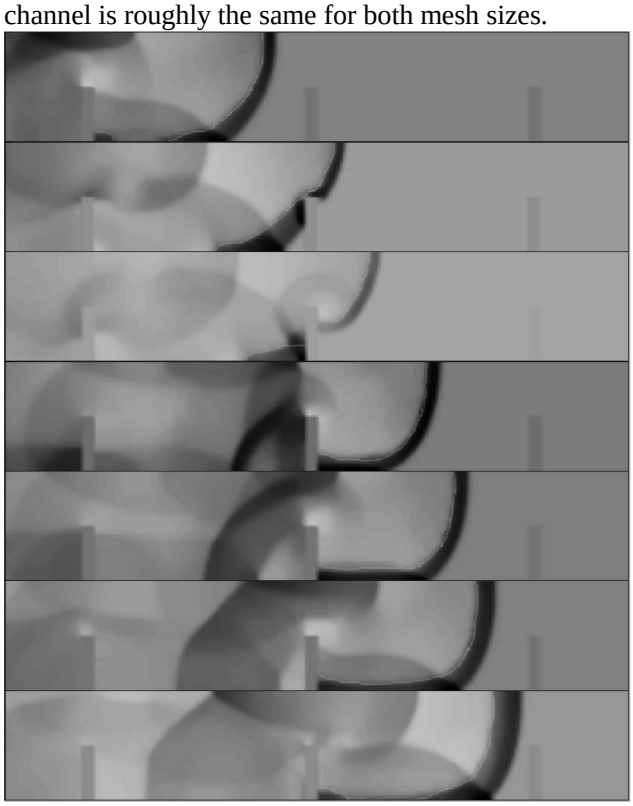


Figure 5: Simulated density gradients of shock-flame-obstacle interactions for the 40 mm channel with repeated obstacles. The images show the failure of detonation. Stoichiometric hydrogen-air at 293 K and 1 atm.

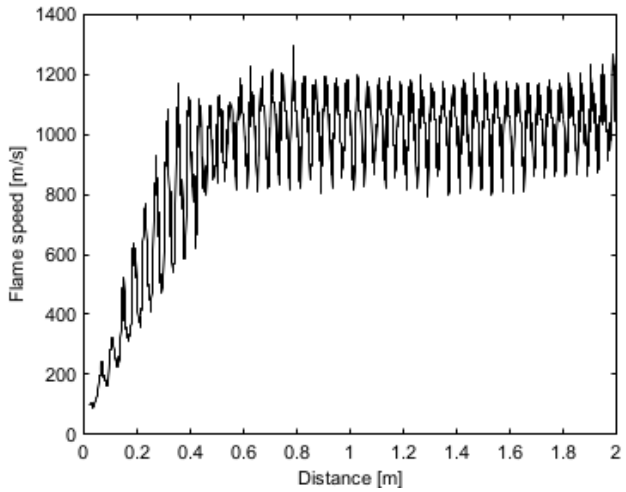


Figure 6: Simulated flame speed as a function of time for the 20 mm channel with repeated obstacles with 1 mm mesh. Stoichiometric hydrogen-air at 293 K and 1 atm.

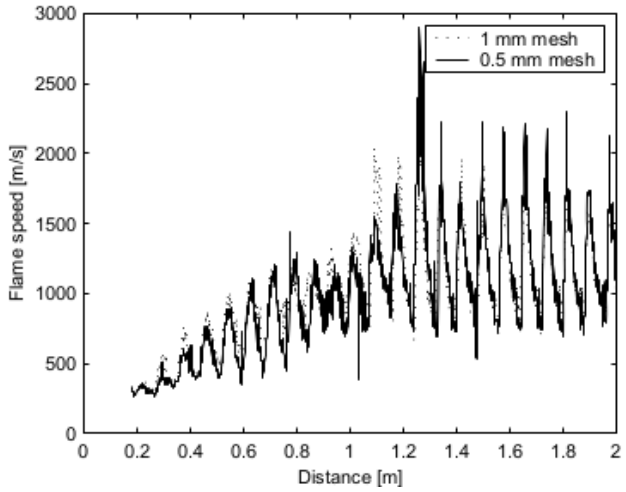


Figure 7: Grid sensitivity for the 40 mm channel with repeated obstacles for 1 mm mesh and 0.5 mm mesh. Stoichiometric hydrogen-air at 293 K and 1 atm.

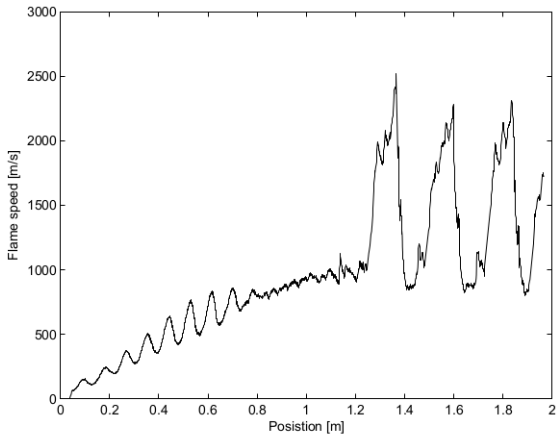


Figure 8: 3D simulation with 1 mm mesh of the flame speed along the center of the channel top wall for the 40 mm channel with repeated obstacles. Stoichiometric hydrogen-air at 293 K and 1 atm.

5 Results and discussion, 3D simulation

Figure 8 shows the simulated flame speed along the channel. Figure 9, Figure 10 and Figure 11 show the experimental pressure histories at 795 mm, 875 mm, and 955 mm from ignition, respectively, as well as the simulated pressure three obstacle distances farther down the channel. The experimental pressure records are extracted from the image files in Teodorczyk (2007) by a simple code. The accuracy of the extraction is not validated but it should reproduce the same curves as in the paper. The simulated time is set to match the strong pressure peak in Figure 9, since this peak is thought to be due to the initiation of the detonation.

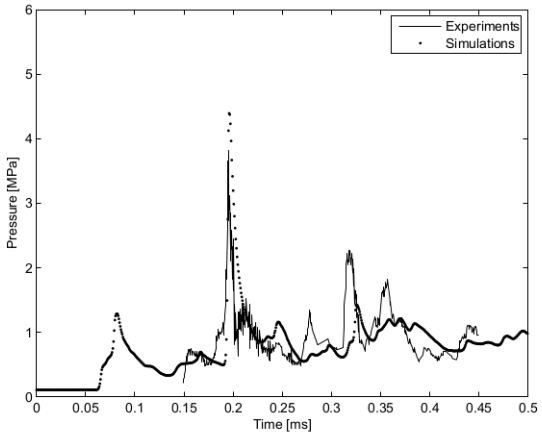


Figure 9: Experimental and simulated pressure history in the 40 mm channel with repeated obstacles. Stoichiometric hydrogen-air at 293 K and 1 atm. The pressure transducer is 795 mm from ignition; the transducer in the simulation is placed three obstacle distances farther from ignition.

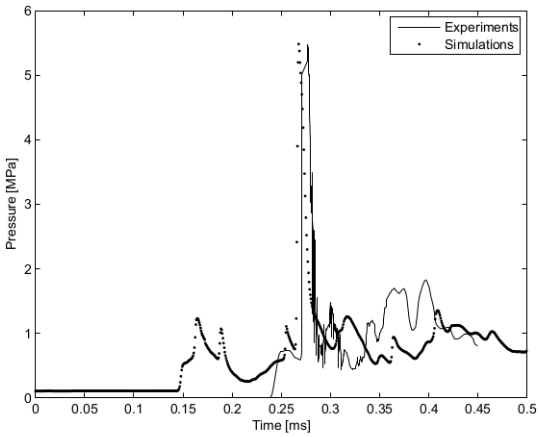


Figure 10: Experimental and simulated pressure history in the 40 mm channel with repeated obstacles. Stoichiometric hydrogen-air at 293 K and 1 atm. The pressure transducer is 875 mm from ignition; the transducer in the simulation is placed three obstacle distances farther from ignition.

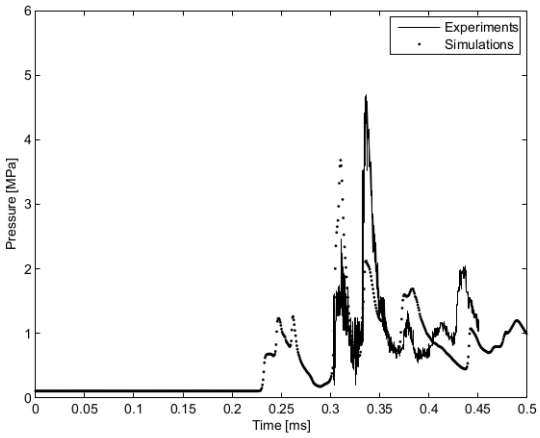


Figure 11: Experimental and simulated pressure history in the 40 mm channel with repeated obstacles. Stoichiometric hydrogen-air at 293 K and 1 atm. The pressure transducer is 955 mm from ignition; the transducer in the simulation is placed three obstacle distances farther from ignition.

The pressure histories from the simulation are taken at a transducer position that is three obstacle distances farther from ignition. Since there is significant scatter in the experiments, the position of the first DDT is difficult to match with the experimental pressure data. The distance between the first pressure rise and the shock from the initiation of the detonation is longer in the simulations than in the experiments. Since the simulated pressure is shifted three obstacle distances, the leading shock has propagated farther from the flame. The coarse resolution of the mesh does not capture all the hotspots. This might average out small areas of high temperature to a lower temperature, whereas the hotspots may ignite and lead to detonation in the experiments. Compared with the 2D simulation, the 3D simulation predicts the first DDT one obstacle later. This might be because any strong shock produced from focusing may propagate in three directions, compared with two directions in the 2D case; furthermore, a simulated hotspot in 2D may lead to transition while in 3D it weakens faster and may not cause transition. Another reason might be that in 2D, the gas is ignited in the entire width of the channel and the flame propagates cylindrically and not spherically, so that the position is moved to where a sufficiently strong shock wave is formed. The 3D simulation does not show the same frequency of the pressure oscillations as with the 2D simulation. The flame propagates with the detonation velocity for the length of three obstacles, compared with only one for the 2D simulation. The shock diffraction is not as critical for the 3D simulation since the propagating detonation front is not planar. As the detonation passes the obstacle, parts of the detonation fail and cause transverse waves that keep the detonation going as a CJ detonation. Previous work on simulation of transition to detonation has been performed with much finer mesh (Gamezo et. al., 2007) but with simpler chemistry. The transition phenomena can be predicted with the present method even when the mesh size is larger than the flame thickness.

6 Conclusions

Simulations of flame acceleration and DDT using the present method reproduce the main effects seen in experiments of flame acceleration in channels with repeated obstacles. Shock focusing and reflections are the most important sources for producing hotspots that lead to the onset of detonation in channels with repeated obstacles. The same processes that led to detonation in the experimental results in Teodorczyk (2007) and Teodorczyk et al. (1988) are seen in the present results. The 3D simulation shows similar behavior to the 2D simulation where DDT occurs, and the flame propagates in the quasi-detonation regime. The simulated initiation and failure of detonation shows that this geometry with a point ignition behaves three-dimensionally, and the details are handled differently for 2D and 3D. The coarse mesh might be the reason for the difference between the simulations and experiments. Since the coarse mesh averages the flame over a few millimeters, the details in the formation of hotspots and the diffraction of the front are not captured. The experiments show significant scatter in the position of the DDT, and it is difficult to say how good the predictions are on that account.

References

G. Ciccarelli, and S. Dorofeev. Flame Acceleration and Transition to Detonation in Ducts. *Progress in Combustion Science*, 34(4): 499-550, 2008.

P. Flohr, and H. Pitsch. A turbulent flame speed closure model for LES of industrial burner flows. *Centre for Turbulent Research Proceedings of the Summer Program*, 2000.

V. N. Gamezo, T. Ogawa, and E. S. Oran. Numerical simulations of flame propagation and DDT in obstructed channels filled with hydrogen-air mixures. In *Proceedings of the Combustion Institute* 31: 2463-2471, 2007.

T. Iijima, and T. Takeno. Effects of temperature and pressure on burning velocity. *Proceedings of the Faculty of Engineering, Tokai Univ*. Vol. X:53-67, 1984.

V. P. Korobeinikov, V. A. Levin, V. V. Markov, and G. G. Chernyi. Propagation of blast in a combustible gas. *Astronautica Acta*, 17: 529–537, 1972.

J. H. S. Lee, and I. O. Moen. The mechanism of transition from deflagration to detonation in vapor cloud explosions. *Progress in Energy Combustion Science* 6: 359-389, 1980.

J. H. Lee, R. Knystautas, and C. K. Chan. Turbulent flame propagation in obstacle filled tubes. In *Proceedings of the Combustion Institute*. 20: 1663-1672, 1985.

J. E. Shepherd, and J. H. S. Lee. On the transition from deflagration to detonation. In: M. Y. Hussaini, A. Kumar and R. G. Voigt, Editors, *Major research topics in Combustion*, 1992.

J. E. Shepherd. Detonation in gases. In *Proceedings of the Combustion Institute* 32: 83-98, 2009.

M. Sichel, N. A. Tonello, E. S. Oran, and D. A. Jones. A two–step kinetics model for numerical simulation of explosions and detonations in H2–O2 mixtures. *Proceedings of the Royal Society A*. 458: 49-82, 2002.

A. Teodorczyk, J. H. S. Lee, and R. Knystautas. Propagation Mechanism of Quasi-Detonations. In *Proceedings of the Combustion Institute* 22: 1723-1731, 1988.

A. Teodorczyk. Scale effects on hydrogen-air fast deflagrations and detonations in small obstructed channels. *Journal of Loss Prevention in the Process Industries*, 21(2): 147-153, 2007. doi:10.1016/j.jlp.2007.06.017.

G. O. Thomas. Some observations on the initiation and onset of detonation. *Philosofical Transactions of the Royal Society A*. 370(1960): 715-739, 2012.

E. F. Toro. *Riemann solvers and Numerical methods for fluid dynamics*. Springer-Verlag Berlin Heidelberg, 1999, (ISBN 3-540-65966-8).

K. Vaagsaether. *Modelling of gas explosions*. PhD Thesis, 2010, Telemark Open Research Archive (http://teora.hit.no/dspace/), http://hdl.handle.net/2282/1113.

Modelling and Simulation of Phase Transition in Compressed Liquefied CO_2

Sindre Tosse Per Morten Hansen Knut Vaagsaether

Department of Process, Energy and Environment Technology, University College of Southeast Norway, Norway,
knutv@usn.no

Abstract

A model and solution method for phase transition in compressed liquefied gases is presented. The model is a two-phase 6-equation model with a common flow velocity for the two phases. The numerical method for solving the model is based on the 2. order shock capturing MUSCL-scheme with a HLLC Riemann solver. The van der Waal cubic equation of state is used for closing the set of equations. The phase transition model is based on thermodynamic and mechanical relaxation between the phases. The goal of the work is to present a numerical model capable of resolving the two-phase flow situation in the depressurization of a vessel or pipe containing liquefied CO_2. Simulation of expansion and phase transition in pressurized liquefied CO_2 is presented and compared with experimental data. The simulations are with a one dimensional geometry and the experiments are performed in a narrow tube. Wall effects in the experiments are not captured in the simulations. The wave structure seen in the experiments is reproduced by the simulation although not quantitatively. The simulations show that the fluid is in the metastable region before it undergoes a phase transition. The level of expansion of the metastable liquid shown in the in the simulations is not seen in the experiments.

Keywords: phase transition, liquefied gas, BLEVE, van der Waal, MUSCL

1 Introduction

The focus of this paper is to present a numerical model capable of resolving the two-phase flow situation in the depressurization of a vessel or pipe containing liquefied CO_2. The methodology is attended for use with all types of liquefied pressurized gases. Sublimation of solid particles will not be addressed, since liquid-vapour interaction is the dominant process inside and in the immediate vicinity of the vessel. In order to get the necessary level of accuracy in the thermodynamic calculations, a non-monotonic equation of state is chosen. For CO_2, the most accurate liquid-vapour EOS available is the Span-Wagner multiparameter EOS (Span and Wagner, 1996). It would be extremely challenging to implement this type of EOS into a numerical code, but the authors regards this as the end-goal of the present work. The usage of a non-monotonic EOS in a numerical solver raises a number of issues, since both the liquid and vapour states have a limited region of existence. In order to deal with these issues, the simplest form of a non-monotonic EOS, namely the cubic van der Waals EOS, is used in the development of a numerical code. Menikoff and Plohr (1989) state that the Maxwell equal-area rule must be applied to modify the equation of state in order to avoid imaginary speed of sound in the van der Waals loop. Saurel et al. (2008) propagate the misconception that the square speed of sound is negative in the spinodal zone. In the present work however, a less strict method is applied to allow metastable states, while maintaining a real speed of sound. While quantitatively inaccurate, the van der Waals equation of state provides a qualitative representation of every major feature of real gas behavior. Combined with its simple formulation, this makes it an often used EOS in model development and academic work. Using a non-monotonic equation of state in a numerical solver raises a number of issues. It is therefore necessary to develop robust solving algorithms that are capable of handling two phase flow in the vicinity of spinodal states. The van der Waals EOS is chosen to develop a proof of concept, because its simple formulation allows for analytical expressions for many thermodynamic parameters, e.g. the spinodal curve. Most compressible two-phase solvers use some form of stiffened gas equation of state or a more generalized Mie Gruneisen form equation of state. Even though it can be written on Mie-Gruneisen form, the van der Waals equation of state has been little used in the context of fluid dynamics. Slemrod (1984) analyzed the dynamic phase transitions in a van der Waals fluid. Zheng et al. (2011) used an interface capturing method with a generalized equation of state on the Mie-Gruneisen form where, among others, the van der Waals equation of state was used. To the authors knowledge, no solvers allowing metastable two-phase compressible flow with phase transition using the van der Waals equation of state exists.

Some work has been done to develop numerical models that are capable of describing evaporation waves. Saurel et al. (1999) developed a Godunov method for compressible multiphase flow that was later applied to the subject of phase transition in metastable liquids (Saurel et al., 2008). They were able to qualitatively reproduce the evaporation front velocities measured by Simoes-Moreira and Shepherd (1999). In recent years, there have been several at-

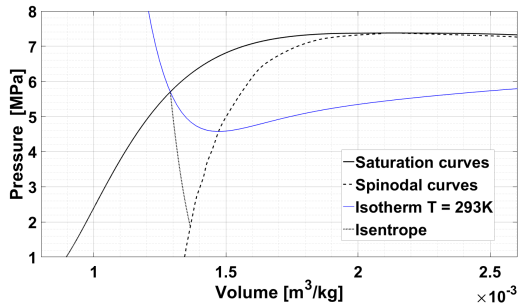


Figure 1. Pressure-volume diagram for CO_2 showing saturation curve, spinodal curve, an isotherm and an isentrope.

tempts to model BLEVE-type scenarios (Pinhasi et al., 2007; VanDerVoort et al., 2012; Xie, 2013).

1.1 Metastable liquids

Figure 1 shows the pressure-volume diagram of CO_2 calculated from the Span-Wagner EOS. The spinodal curve is defined as $\left(\frac{\partial p}{\partial v}\right)_T = 0$ and is seen as an absolute boundary for an expanding liquid state. In the region between the liquid saturation curve and the spinodal curve a metastable liquid can exist. A metastable liquid is not in an equilibrium condition and a fluid can only stay in such a state for very short times. During a rapid expansion of a compressed liquefied gas metastable liquid states will occur behind propagating expansion waves before phase transition forces the thermodynamic state to change towards equilibrium conditions.

2 Model for two phase flow and phase transition

The numerical model used in this work solves the two-pressure 6-equation model given by Saurel et al. (2009). Without heat and mass transfer, the model reads:

$$\frac{\partial \alpha_1}{\partial t} + u\frac{\partial \alpha_1}{\partial x} = \mu(p_1 - p_2), \tag{1}$$

$$\frac{\partial \alpha_1\rho_1}{\partial t} + \frac{\partial \alpha_1\rho_1 u}{\partial x} = 0, \tag{2}$$

$$\frac{\partial \alpha_2\rho_2}{\partial t} + \frac{\partial \alpha_2\rho_2 u}{\partial x} = 0, \tag{3}$$

$$\frac{\partial \rho u}{\partial t} + \frac{\partial \rho u^2 + (\alpha_1 p_1 + \alpha_2 p_2)}{\partial x} = 0, \tag{4}$$

$$\frac{\partial \alpha_1\rho_1 e_1}{\partial t} + \frac{\partial \alpha_1\rho_1 e_1 u}{\partial x} + \alpha_1 p_1 \frac{\partial u}{\partial x} = -p_I\mu(p_1 - p_2), \tag{5}$$

$$\frac{\partial \alpha_2\rho_2 e_2}{\partial t} + \frac{\partial \alpha_2\rho_2 e_2 u}{\partial x} + \alpha_2 p_2 \frac{\partial u}{\partial x} = p_I\mu(p_1 - p_2). \tag{6}$$

The right hand side terms corresponds to pressure relaxation. p_I is the interfacial pressure, estimated by

$$p_I = \frac{Z_2 p_1 + Z_1 p_2}{Z_1 + Z_2}, \tag{7}$$

where $Z_k = \rho_k c_k$ is the acoustic impedance of phase k. Where α_k is the volume fraction of phase k, ρ_k is the density of phase k, p_k is the pressure of phase k, e_k is the specific internal energy of phase k, Y_k is the mass fraction of phase k, c_k is the speed of sound of phase k, μ is the dynamic compaction viscosity and determines the rate of pressure relaxation, u is the flow velocity and an infinitesimal relaxation time, or large enough drag, is assumed leading to a common velocity between the phases. Phase 1 and phase 2 is vapour and liquid respectively. The mixture speed of sound used in this model is the frozen speed of sound,

$$c_f^2 = Y_1 c_1^2 + Y_2 c_2^2. \tag{8}$$

In the present work, we use stiff pressure relaxation ($\mu \to \infty$). As shown in (Saurel et al., 2009), this means the recovery of the 5-equation model. Then the model is strictly hyperbolic with wave speeds $(u + c_f, u - c_f, u)$.

2.1 The van der Waals equation of state

The van der Waals equation of state (vdW-EOS) is the simplest form of a cubic equation of state. It is classified as cubic because it can be written on the form

$$v^3 + a_2 v^2 + a_1 v + a_0 = 0 \tag{9}$$

where v is the specific volume and a_k are pressure and/or temperature dependent coefficients. The vdW-EOS can be derived from the ideal gas EOS by adding correction terms for the excluded volume occupied by finite-sized particles and inter-molecular forces. On its classical form, the vdW-EOS reads

$$\left(p + \frac{n^2 a}{V^2}\right)(V - nb) = nR_M T \tag{10}$$

where n is the number of moles occupying the volume V at pressure p and temperature T. R_M is the ideal gas constant. a is a measure of the attraction between particles and b is the volume excluded by one mole of particles (molecules). As the volume tends to infinity, the vdW-EOS converges to the ideal gas law. The special case of $V = nb$ corresponds to a situation where the volume V is completely filled by the particles. At this point, the pressure tends to infinity. This implies that the van der Waals equation of state is only valid for $V > nb$. In terms of the volume at the critical point, this limit can be written as $\frac{v}{v_c} > \frac{1}{3}$.

3 Solver

The equation set is solved by the 2. order accurate shock capturing MUSCL-scheme (Monotone Upstream-centered Scheme for Conservation Laws) combined with a HLLC (Harten Lax vanLeer Contact) Riemann solver for the interfacial fluxes (Toro, 1999). This solver is used for the hyperbolic part of the equation set i.e. the left hand side of equations 2 to 6.

The shock capturing method with the approximate Riemann solver solves shock waves and contact surfaces as very steep gradients with a numerical diffusion of a shock or contact discontinuity thickness of usually three control volumes. The equation set is closed by the van der Waals equation of state. The time step is variable and controlled by the Courant-Friedrich-Levy number.

3.1 Stiff pressure relaxation

The pressure relaxation step solves the equation set

$$\frac{\partial \alpha_1}{\partial t} = \mu(p_1 - p_2), \tag{11}$$

$$\frac{\partial \alpha_1 \rho_1 e_1}{\partial t} = -p_I \mu(p_1 - p_2), \tag{12}$$

$$\frac{\partial \alpha_2 \rho_2 e_2}{\partial t} = p_I \mu(p_1 - p_2) \tag{13}$$

in the limit $\mu \to \infty$. All other conserved variable groups are held constant during the relaxation step. According to (Saurel et al., 2009), this system of equations can be replaced by

$$e_k(p, v_k) - e_k^0(p_k^0, v_k^0) + \hat{p}_I(v_k - v_k^0) = 0,\ k = 1,2 \tag{14}$$

and the saturation constraint

$$(\alpha\rho)_1 v_1 + (\alpha\rho)_2 v_2 = 1 \tag{15}$$

where $(\alpha\rho)_k$ is constant during the relaxation step. The system can be closed by the van der Waal equation of state $e_k(\rho_k, p_k)$. Equation 14 can then be reformulated to $v_k(p)$ by using an estimate of $\hat{p}_I$. In the present work, the estimation $\hat{p}_I = p_I^0$ is used, but other estimates can also be used as shown by Saurel et al. (2009). Finally, we insert the expressions for v_k into eq. 15 and solve for p.

Since the pressure estimated by this method is not guaranteed to be in agreement with the mixture equation of state $p(\rho, e, \alpha_1)$, this pressure is only used to find the relaxed volume fraction α_1. The relaxed pressure is then determined by the mixture equation of state and the internal energy from the redundant total energy equation. The conserved variables $(\alpha\rho e)_k$ are then re-initialized using the relaxed pressure and volume fraction. This ensures the conservation of mixture energy in the flow field.

Alternate relaxation methods can also be used. Both isentropic and isenthalpic relaxation methods has been tested with the same results as the method described here. This gives reason to assume that the thermodynamic relaxation path is of lesser importance, since it is only used to estimate the relaxed volume fraction. If the numerical method is expanded to a more complex EOS, this means that the pressure relaxation process most likely can be resolved with a less rigorous estimate of the thermodynamic relaxation path.

With the reduced vdW-EOS, eq. 14 can be written as

$$\pi(\delta_k) = \frac{2C_k\delta_k^2\hat{\pi}_I - 2\delta_k^3\hat{\pi}_I - 3\delta_k + 3}{\delta_k^2(3\delta_k - 1)}, \tag{16}$$

where

$$C_k = \delta_k^0 + \frac{1}{\hat{\pi}_I}\left[\frac{1}{2}(\pi_k^0 + \frac{3}{(\delta_k^0)^2})(3\delta_k^0 - 1) - \frac{3}{\delta_k^0}\right]. \tag{17}$$

Since we have no mass transfer, we can write

$$G_1\delta_1 + G_2\delta_2 = 1. \tag{18}$$

where $G_k = (\alpha\rho)_k v_c$. From this, we get

$$\delta_2(\delta_1) = \frac{1 - G_1\delta_1}{G_2} \tag{19}$$

The algorithm for stiff pressure relaxation solves the equation $f(\delta_1) = 0$ by the Newton-Raphson method, where

$$f(\delta_1) = \pi_1(\delta_1) - \pi_2(\delta_1),$$
$$\frac{\mathrm{d}\pi_k}{\mathrm{d}\delta_1} = \left(-\hat{\pi}_I\frac{6C_k - 2}{(3\delta_k - 1)^2} + \frac{6(3\delta_k^2 - 5\delta_k + 1)}{\delta_k^3(3\delta_k - 1)^2}\right)d_k,$$
$$d_1 = 1,\ d_2 = -\frac{G_1}{G_2}$$

Where π is reduced pressure and δ is reduced volume.

3.2 Stiff thermodynamic relaxation

The thermodynamic relaxation method used presently differs somewhat from the methods used by Saurel et al. (2008) and Zein et al. (2010). It is simpler in formulation and relatively easy to implement for any equation of state. We consider a two phase system with total density $\rho = \alpha_1\rho_1 + \alpha_2\rho_2$ and total internal energy $e = Y_1e_1 + Y_2e_2$. Since no mass or heat is added to the system during the relaxation step, these mixture properties are constant. We will consider the velocity of the two phases to be equal and constant during the relaxation step. Initially, the system is closed by the known variables ρ_1, ρ_2, e_1, e_2. In the numerical solver used presently, the two phases will be in mechanical equilibrium at the start of the relaxation step, but this is not a prerequisite of the procedure. The system can be uniquely determined by requiring complete thermodynamic equilibrium between the two phases:

$$p_1 = p_2 = p,\ T_1 = T_2 = T,\ g_1 = g_2 = g. \tag{20}$$

Where g is the Gibbs free energy. Note that this requirement is not possible for all ρ and e. This is indeed the case when there is only a single phase solution, that is when the limit of complete evaporation or condensation is reached. Since the numerical method is only valid for $\alpha_k > \xi$, where ξ is some small number (typically $\xi = 10^{-6}$), the single phase limit of phase 1 will be determined by

$$p_1 = p_2 = p,\ T_1 = T_2 = T,\ \alpha_1 = 1 - \xi \tag{21}$$

and equivalent for the single phase limit of phase 2. If a cubic equation of state is used, even this is not possible for all ρ and e. This will be the case when one phase reaches the spinodal state before thermal equilibrium is reached. If

phase 2 is at the spinodal state, the system is determined by

$$\begin{aligned} p_1 = p_2 = p_{\text{spin}}(v_2) = p, \\ T_1 = T(v_1, p),\ T_2 = T_{\text{spin}}(v_2),\ \alpha_1 = 1 - \xi \end{aligned} \tag{22}$$

v_2 is determined by the mixture equation of state, and v_1 is determined by conservation of mass ($v = Y_1 v_1 + Y_2 v_2$). The subscript *spin* denotes the thermodynamic spinodal state.

In the context of the van der Waals EOS, the three cases (20, 21 and 22) can be identified by the values of ρ and e. A fourth case is theoretically possible, namely $e < e(\rho)_{T=0}$, but this is not likely to occur in numerical calculations and is therefore not further examined.

The stiff thermodynamic relaxation procedure was used when $p_l < p_{\text{sat}}(T_l)$. An additional criterion $\xi_I < \alpha_1 < 1 - \xi_I$ can be used, where ξ_I represents the interface limit of the volume fraction (typically $\xi_I = 10^2\xi$ to $10^3\xi$). This last criterion is referred to as the interface criterion of the thermodynamic relaxation procedure and is used to allow for the formation of metastable liquid.

4 Experiments

The capabilities of the model to predict phase transition in pressurized liquid CO_2 by expansion is validated by comparing simulation results with experimental results. The experimental results are presented in Hansen et al. (2016). Figure 2 shows a drawing of a experimental apparatus for rapid expansion of liquefied CO_2. The expansion tube is 9 mm inner diameter, 1.5 mm wall thickness polycarbonate. Before the beginning of the experiment, the tube is filled to about half level with saturated liquid CO_2 at room temperature, about 20^{o}C. The pressure in the tube is then 5.5 MPa. The top of the tube is closed with a diaphragm which is punctured by an arrow, releasing CO_2 to the atmosphere. Expansion waves then propagates down the tube and starts a boiling process due to the falling pressure. The expansion tube is transparent and a high speed digital camera captures the expansion and boiling process on a high speed movie which is later analyzed. The camera operates at 20 000 fps for this experiment. Typical wave trajectories is shown in figure 3.

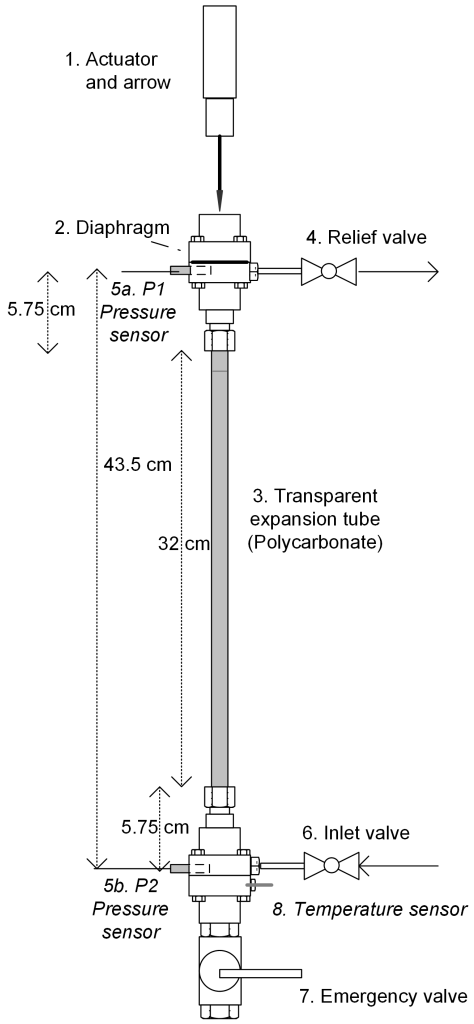


Figure 2. Experimental set-up.

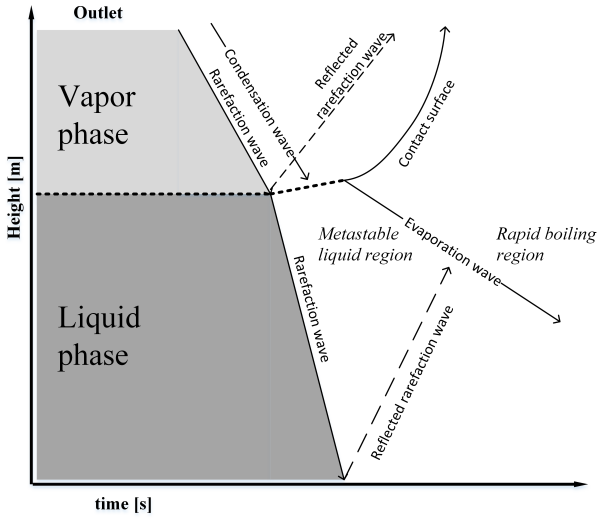


Figure 3. Schematic representation of the waves in the one dimensional expansion experiments.

5 Simulation set-up

The simulation domain is shown with initial conditions in figure 4. The calculation was run with an initial CFL number set to 0.2 for the first 200 time steps. The CFL number was then linearly increased to 0.5 over 50 time steps and was set to 0.5 for the rest of the calculation. The initial conditions for the simulation is shown in table 1. The one dimensional domain was divided into 7000 control volumes with 10^{-4} m length.

6 Results and discussion

The van der Waal EOS is not able to reproduce the thermodynamical states quantitatively, especially close to saturation condition. The results are presented as scaled quantities to show the qualitative behaviour of the simulation method. The pressure is scaled with saturation pressure

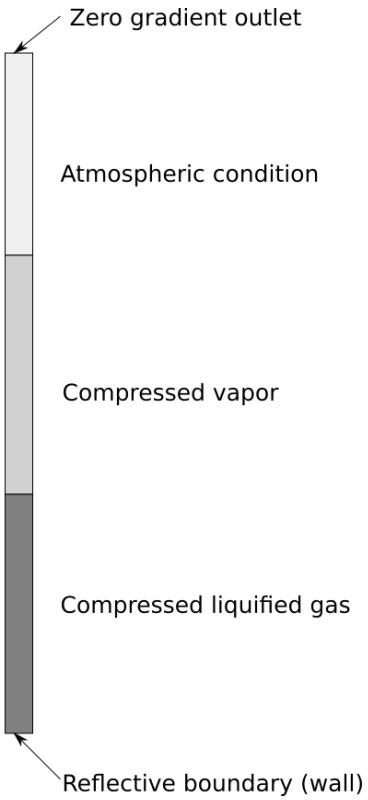


Figure 4. Initial and boundary conditions in the simulation domain.

Table 1. INITIAL SIMULATION CONDITIONS

	$x < 0.25$ m	$0.25m \geq x$ $x < 0.5m$	$x \geq 0.5m$
$p\ [Pa]$	$5.5 \cdot 10^6$	$5.5 \cdot 10^6$	10^5
$u\ [m/s]$	0	0	0
α	10^{-6}	$1-10^{-6}$	$1-10^{-6}$
$\rho_1\ [kg/m^3]$	175.00	175.00	1.8794
$\rho_2\ [kg/m^3]$	530.45	565.46	565.46

at initial temperature, ie. the initial pressure in the tube. The time is scaled by the average propagation time for an expansion wave along the total length of the pipe and the position is scaled by the tube length. The initial interphase in the experiments was 56 % of the tube length from the bottom. For comparison of the wave structures the interphase is moved to scaled position 0.5 like in the simulations. The wave structures in the experiments and simulations are shown as x-t diagrams. The experimental x-t diagram is extracted from the high speed movie. The pixel row from the central position of the tube is stacked along the time vector.

Figure 5 shows the simulated wave structure in the expansion tube. An initial expansion wave propagates downwards in the gas phase from scaled time 0. The expansion wave both reflects and transmits at the interphase, at scaled time 0.7, where the reflected wave is seen traveling upwards and the transmitted wave continues down-

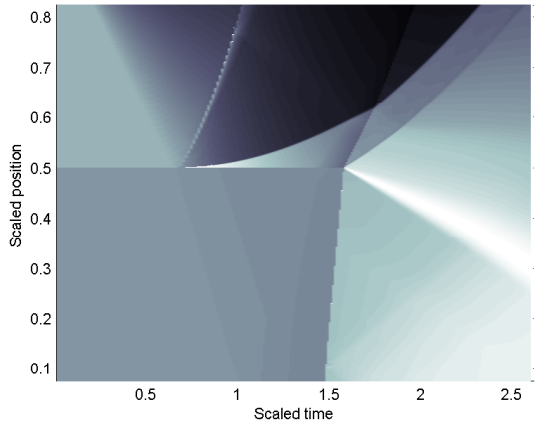


Figure 5. Scaled simulated density for expansion of CO_2 in 1D-domain. The results show the wave structure in the expansion process.

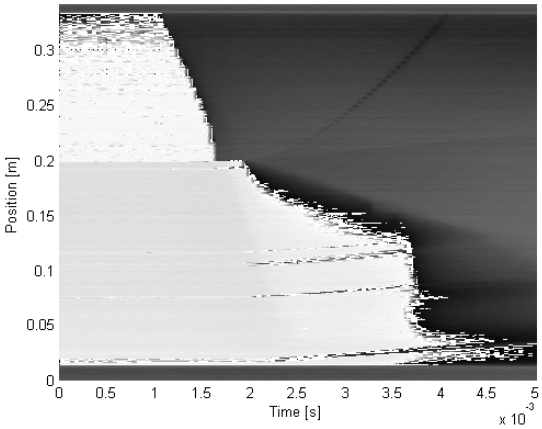


Figure 6. Experimental x-t diagram of expansion of CO_2 in a narrow tube. The results show the wave structure in the expansion process.

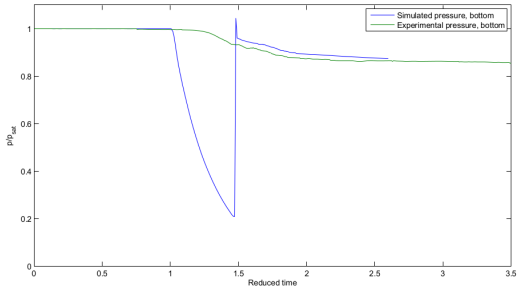


Figure 7. Simulated and experimental scaled pressure history at the bottom of the expansion tube.

wards into the liquid. A condensation phase transition occurs behind the reflected upwards traveling expansion wave. A phase transition in the liquid is initiated and the contact surface of the expanding liquid-gas mixture travels upwards following the reflected expansion wave. The

expansion wave traveling through the liquid is reflected at the bottom of the tube and a faster phase transition is initiated there due to the high level of expansion. The phase transition initiated by the incident expansion wave is slow due to a low level of superheat. Once the expansion wave reflects at the bottom and again interactes with the initial interphase between liquid and vapour, at scaled time 1.6, a faster phase transition is triggered. Comparing these results to the experimental results seen in figure 6 shows the same wave structures. In the experiments a condensation wave following the incident expansion wave occurs. This is not seen in the simulations. The reflected expansion wave is not clearly in seen in the experimental x-t diagram. The condensation seen in the simulations will not occur in experiments since the wave propagates into a two phase fluid.

Figure 7 shows the relative scaled pressure at the bottom of the tube vs. scaled time for simulation and experiment. The large drop in the simulated pressure, not seen in the experiments, is due to the expanding liquid. The thermodynamical state in the expansion wave is highly expanded metastable liquid. When the liquid pressure reaches the spinodal state at scaled time 1.5, a very rapid phase transition occurs and brings the pressure up towards equilibrium pressure. This creates a shock wave propagating upwards due to the fast expansion in the boiling. This shock is driven by a sudden change in thermodynamic state to equilibrium. This rapid phase transition propagates with the mesh speed, ie. $\Delta x/\Delta t$ and is an artefact of the phase transition model. The experimental pressure values does not drop as dramatically as the simulated pressure. The reason for this discrepancy can be that nucleation sites along the narrow tube will force a faster phase transition in the metastable liquid and keep the pressure at a higher level. The wall effects are not included in the simulation. After the rapid phase transition and formation of the shock wave the simulated pressure is close to the experimental pressure.

7 Conclusions

A model and solver for rapid phase transition in compressed liquefied gases is presented. The phase transition model uses a mechanical and thermodynamical relaxation approach for phase transition. The present model and solver is capable of handling the wave types that can occur in a depressurization process however the combination of the van der Waals equation of state and an ideal geometry in one dimension will not produce the quantitative values seen in the experiments. Wall effects and low accuracy of the EOS close to saturation conditions and in metastable state causes a higher degree of superheat before a rapid phase transition can occur in the simulations. When the metastable liquid reaches the spinodal state, the model produces an unphysically fast evaporation wave. Future work to improve the simulation method will be to develop a kinetic based phase transition model in highly expanded metastable liquids. Such a model can reduce the possibility of low pressures seen in the metastable liquid during the reflection of rarefaction waves. A kinetic based transition rate can include wall effects and effects from impurities in the liquid. For higher accuracy the present method can be extended to more complex equations of states, like the Span-Wagner EOS.

References

P. M. Hansen, K. Vaagsaether, A. V. Gaathaug and D. Bjerketvedt. CO_2 explosions - an experimental study of rapid phase transition. *8th International Seminar on Fire and Explosion Hazards*, 25. – 28. April, Hefei, China, 2016.

R. Menikoff and B.J. Plohr. The Riemann Problem for Fluid-Flow of Real Materials. *Reviews of Modern Physics*, 61(1):75–130, 1989.

G. A. Pinhasi, A. Ullmann, and A. Dayan. 1D plane numerical model for boiling liquid expanding vapor explosion (BLEVE). *International Journal of Heat and Mass Transfer*, 50(23-24):4780–4795, 2007.

R. Saurel, and R. Abgrall. A multiphase Godunov method for compressible multifluid and multiphase flows. *Journal of Computational Physics*, 150(2):425–467, 1999.

R. Saurel, F. Petitpas, and R. Abgrall. Modelling phase transition in metastable liquids: application to cavitating and flashing flows. *Journal of Fluid Mechanics*, 607:313–350, 2008.

R. Saurel, F. Petitpas, and R. A. Berry. Simple and efficient relaxation methods for interfaces separating compressible fluids, cavitating flows and shocks in multiphase mixtures. *Journal of Computational Physics*, 228(5):1678–1712, 2009.

J.R. Simoes-Moreira and J.E. Shepherd. Evaporation waves in superheated dodecane. *Journal of Fluid Mechanics*, 382:63–86, 1999.

M. Slemrod. Dynamic phase transitions in a van der Waals fluid. *Journal of Differential Equations*, 52(1):1–23, 1984.

R. Span and W. Wagner. A new equation of state for carbon dioxide covering the fluid region from the triple-point temperature to 1100 K at pressures up to 800 MPa. *Journal of Physical and Chemical Reference Data*, 25(6):1509–1596, 1996.

E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics.* Springer-Verlag, Heidelberg, second edition, 1999.

M. M. Voort, A. C. Berg, D. J. E. M. Roekaerts, M. Xie, and P. C. J. Bruijn. Blast from explosive evaporation

of carbon dioxide: experiment, modeling and physics. *Shock Waves*, 22(2):129–140, 2012.

M. Xie. "*Thermodynamic and Gasdynamic Aspects of a Boiling Liquid Expanding Vapour Explosion.*" *PhD thesis, Delft University of Technology*, 2013.

A. Zein, M. Hantke, and G. Warnecke. Modeling phase transition for compressible two-phase flows applied to metastable liquids. *Journal of Computational Physics*, 229(8):2964–2998, 2010.

H. W. Zheng, C. Shu, Y. T. Chew, and N. Qin. A solution adaptive simulation of compressible multi-fluid flows with general equation of state. *International Journal for Numerical Methods in Fluids*, 67(5):616–637, 2011.

Parallel Simulation of PDE-based Modelica Models using ParModelica

Gustaf Thorslund[1] Mahder Gebremedhin[2] Peter Fritzson[2] Adrian Pop[2]

[1]ThorslundTech AB, Sweden, gustaf@thorslundtech.se
[2]Dept. of Computer and Information Science, Linköping University, Sweden,
{mahder.gebremedhin,peter.fritzson,adrian.pop}@liu.se

Abstract

The Modelica language is a modelling and programming language for modelling cyber-physical systems using equations and algorithms. In this thesis two suggested extensions of the Modelica language are covered. Those are Partial Differential Equations (PDE) and explicit parallelism in algorithmic code. While PDEs are not yet supported by the Modelica language, this article presents a framework for solving PDEs using the algorithmic part of the Modelica language, including parallel extensions. Different numerical solvers have been implemented using the explicit parallel constructs suggested for Modelica by the ParModelica language extensions, and implemented as part of OpenModelica. The solvers have been evaluated using different models, and it can be seen how bigger models are suitable for a parallel solver. The intention has been to write a framework suitable for modelling and parallel simulation of PDEs. This work can, however, also be seen as a case study of how to write a custom solver using parallel algorithmic Modelica and how to evaluate the performance of a parallel solver.

Keywords: OpenModelica, ParModelica, PDE, parallel computing, GPU, GPGPU

1 Introduction

To understand the behavior of a system, it is desirable to write down known relations of the system as equations. Together the equations will form a model of the system. If the equations contain derivatives with respect to one variable, they describe an Ordinary Differential Equation (ODE) or Differential Algebraic Equation (DAE). If, however, the equations contain derivatives with respect to more than one variable, they describe a Partial Differential Equation (PDE).

Modelica[1] is an object oriented language[2] for modeling complex physical systems using equations. The model can then be simulated using a numerical solver. However, Modelica does not currently support modeling partial differential equations. There are suggested extensions for PDEs in (Fritzson, 2014; Saldamli, 2006).

OpenModelica[3] is an open source[4] implementation of the Modelica language, and an active research area.

Given that a PDE can describe a model in several dimensions, the required computations can grow exponentially with the size of the model. This should make it suitable for parallel computing.

1.1 ParModelica

ParModelica (Gebremedhin et al., 2012), implement a suggested extension for explicit parallelism in the algorithmic subset of Modelica. Similar to CUDA and OpenCL, it adds the concept of parallel computation device, device memory, and functions to be called on the device and within the device.

1.2 Previous Research on PDEs in Modelica

An extensive work on PDEs within Modelica has been done (Saldamli, 2006), suggesting language extensions to the Modelica language to support fields and describing spatial domains. Those extensions were implemented in PDEModelica. Unfortunately, PDEModelica has not been maintained during the development of OpenModelica. However, the work is, nevertheless, a good reference for further work.

1.3 Partial Differential Equations (PDE)

A PDE also depends on derivatives with respect to other variables than time. For example coordinates in space, also known as spatial derivatives.

$$\rho_l \frac{\partial^2 \xi(x,t)}{\partial t^2} = F \frac{\partial^2 \xi(x,t)}{\partial x^2} + f_y(x) \tag{1}$$

$$\frac{\partial T}{\partial t} = \kappa \nabla^2 T + \left(\frac{\kappa h}{\lambda}\right) = \kappa \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) + \left(\frac{\kappa h}{\lambda}\right) \tag{2}$$

Equation (1) describes the vibration of a string with constant tension and (2) describes heat conduction, both equations are from (Nordling and Österman, 2006).

[1]http://www.modelica.org/ accessed May 2016

[2]The Modelica language is an open standard and can be downloaded for free. There is also a book (Fritzson, 2014) available with many examples of how to use the language.

[3]http://www.openmodelica.org/ accessed May 2016

[4]http://opensource.org/ accessed May 2016

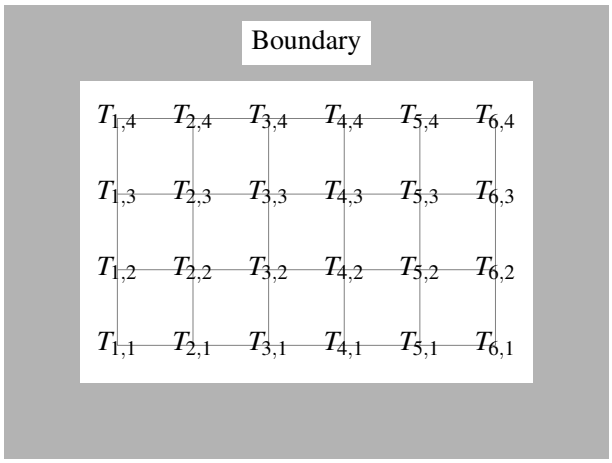


Figure 1. Method of lines applied to the heat conduction equation over a plane.

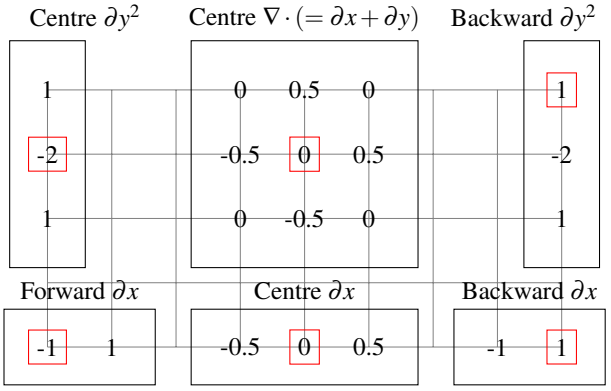


Figure 2. Example of stencils used for calculating spatial derivatives. The red box symbolizes the destination, while the numbers are the weights to use when summing up the neighboring values. They are all approximations, and some can be derived in different ways, resulting in different weights.

1.4 Explicit Form

In control theory and modeling it is often desirable to put an equation into explicit state form, see (Fritzson, 2014; Glad and Ljung, 1989; Ljung and Glad, 2004). In the general form we have the state vector $\vec{x}(t)$, the state derivative $\dot{\vec{x}}(t)$, the input vector $\vec{u}(t)$, and the output vector $\vec{y}(t)$. In the general case we have the equations:

$$\dot{\vec{x}}(t) = \vec{f}(\vec{x}(t),\vec{u}(t)) \tag{3a}$$
$$\vec{y}(t) = \vec{g}(\vec{x}(t),\vec{u}(t)) \tag{3b}$$

In case f and g are linear, matrix notation can be used instead:

$$\dot{\vec{x}}(t) = A\vec{x}(t) + B\vec{u}(t) \tag{4a}$$
$$\vec{y}(t) = C\vec{x}(t) + D\vec{u}(t) \tag{4b}$$

This article will only use the general form in (3). Models where an explicit form cannot be derived will require solver methods not covered here.

2 Numerics

To give a better understanding of the implementation, this section covers the algorithms involved in simulating mathematical models. For further reading, see: (Eldén and Wittmeyer-Koch, 1996; Fritzson, 2014; Ljung and Glad, 2004), or another book covering numerical analysis or applications of numerical analysis.

2.1 Discretisation

To be able to solve a PDE over space and time, one approach is to discretized the PDE over space and this way get a system of ODEs. If we take the heat conduction equation from (Nordling and Österman, 2006), with ∇^2 expanded to two dimensions, and calculate it at $n_x \times n_y$ discrete points in space we get:

$$\begin{aligned}\frac{\partial T_{i,j}}{\partial t} &= \kappa_{i,j}\nabla^2_{i,j}T + \left(\frac{\kappa h}{\lambda}\right)_{i,j} \\ &= \kappa_{i,j}\left(\frac{\partial^2 T_{i,j}}{\partial x^2} + \frac{\partial^2 T_{i,j}}{\partial y^2}\right) + \left(\frac{\kappa h}{\lambda}\right)_{i,j}\end{aligned} \tag{5}$$

Figure 1 shows how T has been discretised over a grid with 6×4 points. The derivatives in (2) can be approximated with:

$$\frac{\partial^2 T}{\partial x^2} = \frac{T_{i+1,j} - 2T_{i,j} + T_{i-1,j}}{\Delta x^2} \tag{6a}$$
$$\frac{\partial^2 T}{\partial y^2} = \frac{T_{i,j+1} - 2T_{i,j} + T_{i,j-1}}{\Delta y^2} \tag{6b}$$

Those approximations can be derived using Taylor series, see for example (Eldén and Wittmeyer-Koch, 1996; Åström, 2015). As seen in (6), the discretisation (in one direction) will depend on the points on both sides. This is called a central difference, while there are also forward and backward differences depending only on points at one side. The weights to used to approximate the spatial derivatives at a given point is commonly referred to as stencils. Different types of stencils are illustrated in Figure 2.

Due to the dependency of points at the sides, the boundaries need to be treated specially. How they are treated depends on the boundary condition of the model. In the heat conduction case one may assume the temperature is constant at the borders, so for example $T_{0,j} = T_{1,j}$, and expand the values at the boundaries. Other models may have other boundary conditions.

Due to the amount of points, with one ODE at each point, the method of lines approach will produce, this can result in fairly large matrices if using an implicit solver. If, on the other hand, an explicit solver is used, this gives a potential for lots of parallelism. When running on a

General Purpose Graphic Processing Unit (GPGPU) each thread can have its own point.

2.2 Runge-Kutta with Variable Step Length

If the value of x_{n+1} is approximated with different order of error, the values can be compared to get an estimate of the local error. In (Bogacki and Shampine, 1989), parameters for calculating both a third and second order approximation using four computations of k are suggested:

$$k_1 = f(x_n, x_n) \tag{7a}$$
$$k_2 = f(t_n + \frac{1}{2}h, x_n + \frac{1}{2}hk_1) \tag{7b}$$
$$k_3 = f(t_n + \frac{3}{4}h, x_n + \frac{3}{4}hk_2) \tag{7c}$$
$$x_{n+1}^{(3)} = x_n + (\frac{2}{9}k_1 + \frac{1}{3}k_2 + \frac{4}{9}k_3)h \tag{7d}$$
$$k_4 = f(t_n + h, x_{n+1}) \tag{7e}$$
$$x_{n+1}^{(2)} = x_n + (\frac{7}{24}k_1 + \frac{1}{4}k_2 + \frac{1}{3}k_3 + \frac{1}{8}k_4)h \tag{7f}$$

Using the two predictions $x_{n+1}^{(3)}$ and $x_{n+1}^{(2)}$ of third and second order, it is possible to estimate the error during the step. The error can be used to decide if the step should be accepted or restarted with a shorter step size. It is also possible to estimate a new step size.

3 General-Purpose Computing on Graphics Processing Units (GPGPU)

A Graphic Processing Unit (GPU) can be used as a computation device attached to a host, Figure 3. Within a GPU there are multiple Computation Unit (CU). The CUs are simplified compared to a CPU, so it is the amount of them that makes the GPU powerful. The GPU will have its own memory, divided into a bigger global memory, and a smaller and faster local memory. The local memory can be used as a user controlled cache. GPUs usually has its own cache too, giving a transparent memory hierarchy. When a GPU device is used within a host computer, it will result in a heterogeneous system. The host can either be used just to control the device, or carry out its own computations.

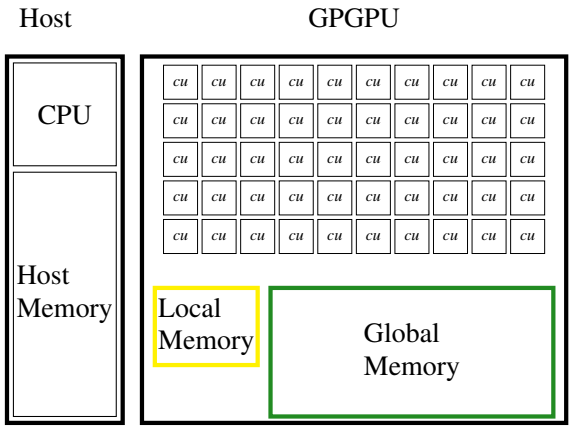


Figure 3. Computer equipped with a GPU. The GPGPU has a number of Computation Units, local and global memory.

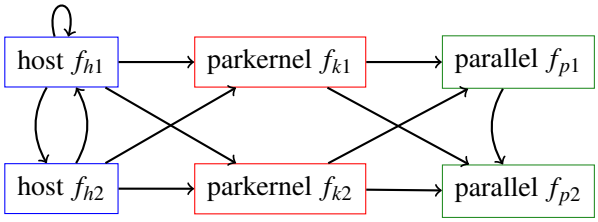


Figure 4. How functions can call each other in ParModelica

4 PDEs in Modelica

PDEs, in general, are not supported in Modelica. Starting with Modelica 3.3 there is support for spatialDistribution, allowing modelling of variable speed transport (Fritzson, 2014). The suggested extension in (Fritzson, 2014; Saldamli, 2006) are:

- field variables
- indomain construct

Those extensions would allow modelling a heat equation on a plane as:

```
model HeatInPlane
  parameter Real c;
  parameter Real q;
  parameter Real h;
  field Real T(domain=omega);
equation
  c*der(T) = pder(T,D.x2) + pder(T,D.y2)
    indomain omage.interior;
  T = 50          indomain omega.left;
  c*pder(T,D.x) = q+h*(T_ext-T)
    indomain omega.right;
  pder(T,D.y) = 0 indomain omega.top;
  pder(T,D.y) = 0 indomain omega.bottom;
end HeatInPlane;
```

5 Algorithmic Modelica and ParModelica

In an algorithmic context, the following concepts from Modelica is of interest:

- Functions
- Records
- Arrays
- Enumeration types
- Type definitions
- Partial function, if implemented by the user they can be passed as argument to an other function

While it is possible to define arrays of arrays, this should be seen as a multidimensional array with rectangular/box-shape. It is not possible to construct arrays containing arrays of different sizes. The same applies if an array of records containing arrays is constructed.
ParModelica adds the concept of:

- *parkernel* function, called from non parallel context
- *parfor* loop, called from non parallel context
- *parallel* function, called from a *parkernel* function or *parfor* loop
- *parlocal* memory, static size set at compile time and used within a *parkernel* function or *parallel* function
- *parglobal* memory, allocated in a non parallel context and passed to a *parkernel* function during executions

Modelica functions can be called recursive. The *parallel* functions introduced by ParModelica can, however, only be called from a parallel context. This is from within a *parkernel* function, *parfor* loop or an other *parallel* function. Furthermore, the *parallel* functions cannot be called recursively. How functions can call each other is illustrated in Figure 4.
ParModelica does, however, add limitations:

- Does not support records
- Does not support partial functions as argument to other functions
- Arrays of arrays should be considered as multidimensional arrays, so it is not possible to define an array on the host containing a number of *parglobal* arrays to be passed to a *parkernel* function

6 Solver Framework

This section gives an overview of the framework for solving PDEs.

6.1 User Defined State Derivative and Settings

The user should provide a function for computing the state derivative. For a solver using ParModelica this should be a parallel function and named ParDerState. This function should be within the PDESolver hierarchy, in the subpackage Model. A serial solver using algorithmic Modelica will instead use the function DerState within same package. The function gets the current state, user provided variables, external fields, a time to compute the state derivative at, and the discrete coordinates as three scalars. The function will then return up to three scalars for up to three different fields. Here the function interface together with a sample model is given:

```
within PDESolver.Model;

parallel function ParDerState
  "Calculate the state derivative"
  import Functions = PDESolver.ParFunctions;
  import PDESolver.ParFunctions.Pder;
  import PDESolver.Types;

  input Types.Field[:] state "Array of state
      fields";
  input Real var[:];
  input Types.Field ext[:];
  input Real t
    "Time to calculate the state derivative
       at";
  input Integer i,j,k
    "Discrete coordinate within field";
  output Real value1;
  output Real value2;
  output Real value3;
protected
  // User defined
  Real d2Tdx2, d2Tdy2;
  Real c = var[1];
algorithm
  // User defined
  nDer := 0; // Perfect insulation
  d2Tdx2 := Pder.Pder2Neumann(f=state, fi=1,
                              i=i, j=j, k=k,
                              dim=1, nder=
                                  nDer);
  d2Tdy2 := Pder.Pder2Neumann(f=state, fi=1,
                              i=i, j=j, k=k,
                              dim=2, nder=
                                  nDer);
  value1 := c*(d2Tdx2 + d2Tdy2)*ext[1,i,j,k];
end ParDerState;
```

For describing the domain and how the field is discretised, the user should provide a Settings package for the model. Here it is also possible to add static parameters for the model.

```
within PDESolver.Model;

package Settings
  // Parameters used by the solver
  constant Types.FieldIndex n = {80,40,1};
  constant Types.Coordinate first = {0,0,0};
  constant Types.Coordinate last = {2,1,0};
  constant Integer stateFields = 1 "I";
```

```
  // Parameters used in the model
  constant Integer boxSize = integer(n[2]/2);
  constant Integer myBoundary = 3;
end Settings;
```

6.2 Types Used by the Solver

There are two types used by the solver field type, implemented as a four dimensional array, and an enumeration type to select solver. For the field type the dimensions are taken from the user provided settings.

```
within PDESolver;

type Field = Real[Model.Settings.n[1],
                  Model.Settings.n[2],
                  Model.Settings.n[3]]
  (each start=0, each fixed=true);
```

```
within PDESolver.Solver;

type SolverId =
  enumeration(SerialPECE,
              SerialRK32,
              ParallelPECE,
              ParallelRK32);
```

6.3 Solvers

The user interface for simulating a PDE is the Solve function. It will take a state array, external fields, user provided variables, the current time, time to step forward to, and a SolverId as compulsory arguments. It is also possible to provide arguments for initial intermediate step size, maximum error during one step, maximum intermediate step size. The arrays provided are host variables and for the parallel solvers they will be copied to *parglobal* variables before calling the solver. The result is then copied back and returned as next state.

```
within PDESolver.Solver;

function Solve
  input Types.Field state[:] "Current state";
  input Types.Field ext[:] "External field";
  input Real var[:] "External variables";
  input Real t0 "Time at state";
  input Real t1 "Time at next";
  input SolverId solverId;
  input Real dt = (t1-t0)*2
    "Initial intermediate step length.";
  input Real eMax = 0.1 "Max error";
  input Real hMax = dt
    "Max intermediate step lengh.";
  input Integer th1=1, th2=1, th3=1;
  output Types.Field next[size(state,1)]
    "New state";
protected
  // ...
algorithm
  // ...
end Solve;
```

The solvers need different intermediate fields. Since a *parkernel* function in ParModelica cannot have internal arrays, output variables are used as intermediate fields.

For merging fields within the parallel solvers, ordinary for-loops are used. Each thread will calculate the initial index, step size, and final value for the loops.

7 Use Case — Heat in Plane

In this section two use cases with heat conduction in a plane and different boundary conditions are presented, together with results and a discussion about result.

7.1 Poor Insulation and Constant Temperature

The boundary conditions here are similar to those in (Fritzson, 2014), with constant temperature at one side, poor insulation at opposite side and perfect insulation at the remaining two sides. Figure 5 shows how the temperature falls from the side with constant temperature to the side with poor insulation, while the sides with perfect insulation does not have an impact on the temperature.

```
if Functions.AtBoundary(i,j,k) then
  if Functions.AtFirst(i=i,j=j,k=k,dim=1)
      then
    // Left side
    d2Tdx2 :=
      Pder.Pder2Dirichlet(f=state, fi=1,
                          i=i, j=j, k=k,
                          dim=1, boundary=50)
                              ;
  elseif Functions.AtLast(i=i,j=j,k=k,dim=1)
      then
    // Right side
    nDer := q + h*(T_ext - state[1,i,j,1]);
    d2Tdx2 :=
      Pder.Pder2Neumann(f=state, fi=1,
                        i=i, j=j, k=k,
                        dim=1, nder=nDer);
  else
    d2Tdy2 :=
      Pder.Pder2Neumann(f=state, fi=1,
                        i=i, j=j, k=k,
                        dim=2, nder=0);
  end if;
else
  d2Tdx2 :=
    Pder.Pder2Inner(f=state,
                    fi=1, i=i, j=j, k=k, dim=
                        1);
  d2Tdy2 :=
    Pder.Pder2Inner(f=state,
                    fi=1, i=i, j=j, k=k, dim=
                        2);
end if;
value1 := c*(d2Tdx2 + d2Tdy2);
```

7.2 Constant Temperature Depending on Location

In this example boundaries have constant temperature depending on location around the plate. The result after a 500ms simulation can be seen in Figure 6. The tempera-

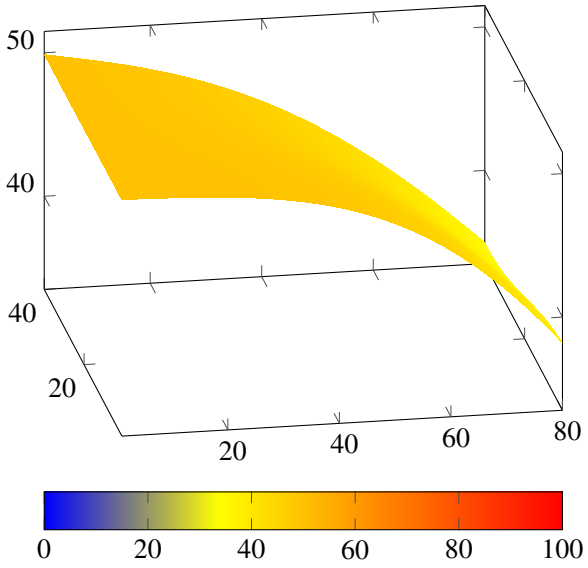


Figure 5. Plane with constant heat at the left side and poor insulation at the right side ($t = 0.5$).

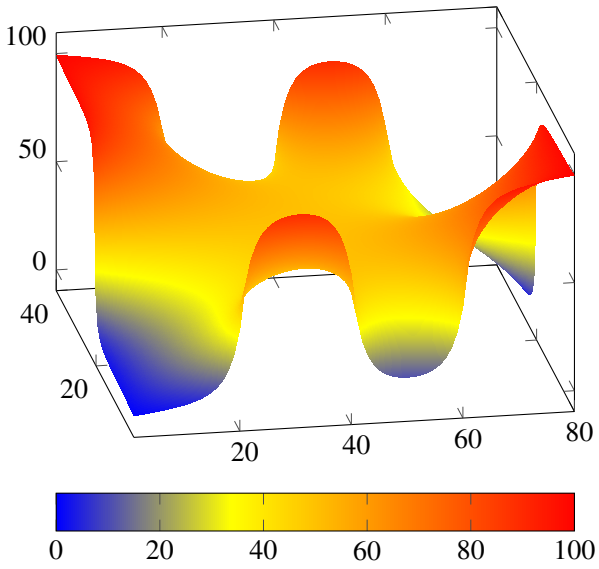


Figure 6. Boundary have constant temperature depending on location ($t = 0.5$).

ture have been forced towards 0 or 100 depending on the boundary conditions.

```
boundaryT := 100*BoxZeroOne(i,j);
d2Tdx2 :=
  Pder.Pder2Dirichlet(f=state, fi=1,
                      i=i, j=j, k=k,
                      dim=1, boundary=
                          boundaryT);
d2Tdy2 :=
  Pder.Pder2Dirichlet(f=state, fi=1,
                      i=i, j=j, k=k,
                      dim=2, boundary=
                          boundaryT);
value1 := c*(d2Tdx2 + d2Tdy2);
```

8 Performance Measurement

Performance measurement where done on a Fermi M2050 GPU with a total of 448 CUDA cores available. The same simulation was started using different number of threads and the speedup compared to just using one thread can be seen in Figure 7. In parallel computation there will always be a sequential part limiting the maximum speedup. This is due to the sequential part will take the same amount of time no matter how fast the parallel part may run.

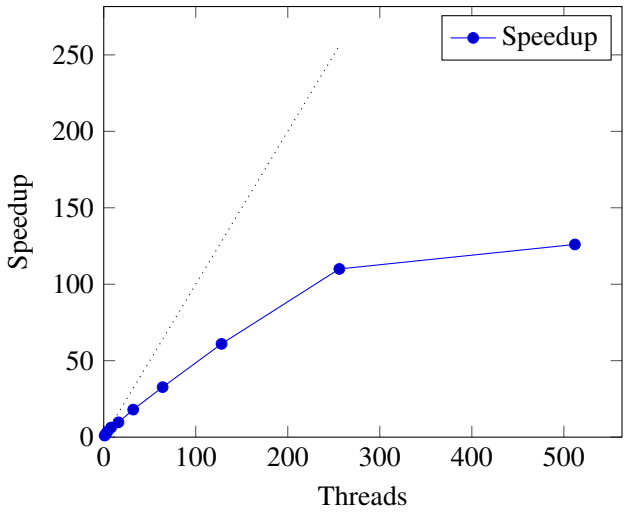


Figure 7. Speedup when simulating same model using different amount of threads for the simulation on a GPU with 448 CUDA cores.

9 Pros & Cons of Solver Written in Modelica

In the context of a bigger system, where part of it need a special solver, there can be advantages to write the solver in Modelica. The solver can then form a framework where the user only add a small portion of code. For this the suggested ParModelica extensions may also be used to gain better performance. Another advantage with ParModelica can be when evaluating the performance of a potential parallel solver. Then the solver can be written using ParModelica to get an idea of performance gain and bottlenecks when appying the solver to different problems.

The true power of Modelica is to solve equations. While the language does have an algorithmic subset, it is hard to compete with other general purpose programming languages.

10 Conclusions

In this article we show how an algorithmic solver can be implemented to utilise the explicit parallelism of ParModelica. To gain performance, however, care must be taken of where the user code is added. If the solver is written as a kernel function, calling a parallel function provided by the user, it is possible to get two order of magnitudes

better performance.

While this work tries to provide the functionality to solve various PDEs, it is hard to predict all needed functionality without specific use cases. For this it is necessary to simulate more models. Stability and errors introduced by the discretisation and solvers also need further work. There is also ongoing work, by PhD student Jan Šilar, to add PDE extensions to the frontend. This was presented during the OpenModelica workshop 2016, (Šilar, 2016). Once OpenModelica have PDE extensions in the frontend, and there is an efficient way to simulate PDEs, this needs to be integrated into all stages of the OpenModelica compiler and simulation runtime.

This work is still a research prototype and not stable enough to be included in the OpenModelica release. The work we present was initiated by (Thorslund, 2015).

References

P. Bogacki and L.F. Shampine. A 3(2) pair of Runge - Kutta formulas. *Appl. Math. Lett.*, 2(4):321–325, 1989.

Lars Eldén and Linde Wittmeyer-Koch. *Numerisk analys — en intruduktion*. Studentlitteratur, third edition, 1996.

Peter Fritzson. *Principles of Object-Oriented Modeling and Simulation with Modelica 3.3. A cyber-physical approach*. Wiley, second edition, 2014. ISBN 9781118859124.

Mahder Gebremedhin, Afshin Hemmati Moghadam, Peter Fritzson, and Kristian Stavåker. A data-parallel algorithmic modelica extension for efficient execution on multi-core platforms. In *Proceedings of the 9th International MODELICA Conference; September 3-5; 2012;*, pages 393–404, Munich, Germany, September 2012.

Torkel Glad and Lennart Ljung. *Reglerteknik Grundläggande teori*. Studentlitteratur, second edition, 1989.

Lennart Ljung and Torkel Glad. *Modellbygge och simulering*. Studentlitteratur, second edition, 2004.

Carl Nordling and Jonny Österman. *Physics Handbook*. Studentlitteratur, 2006. ISBN 978-91-44-04453-8.

Levon Saldamli. *PDEModelica – A High-Level Language for Modeling with Partial Differential Equations*. PhD thesis, Linköping University, PELAB - Programming Environment Laboratory, The Institute of Technology, 2006.

Gustaf Thorslund. Simulating partial differential equations using the explicit parallelism of ParModelica. Master's thesis, Linköping University, Software and Systems, Faculty of Science & Engineering, 2015.

Freddie Åström. *Variational Tensor-Based Models for Image Diffusion in Non-Linear Domains*. PhD thesis, Department of Electrical Engineering, Linköping University, 2015.

Jan Šilar. Partial Differential Equations in Modelica. OpenModelica2016-talk12-JanSilar-PartialDifferentialEquationsinModelica.pdf, 2016. URL `https://openmodelica.org/events/openmodelica-workshop/openmodelica-program-2016`.

Blood Flow in the Abdominal Aorta Post 'Chimney' Endovascular Aneurysm Repair

Hila Ben Gur [1] Moshe Halak [2] Moshe Brand [3]

[1]School of Mechanical Engineering Faculty of Engineering, Tel Aviv University Tel Aviv, Israel,
hilabengur@mail.tau.ac.il

[3]Department of Vascular Surgery, The Chaim Sheba Medical Center, Tel Hashomer, Israel
moshe.halak@sheba.health.gov.il

[3]Department of Mechanical Engineering and Mechatronics, Faculty of Engineering, Ariel University, Ariel, Israel
mosheb@ariel.ac.il

Abstract

Aortic aneurysms are a main death cause in the elderly population throughout the western world. In recent years, more aneurysm repairs are performed endovascularly using stent grafts (SGs) inserted into the aneurysm site through the arterial system (minimally invasive). In this study, we analyze the hemodynamics in an aneurysmatic abdominal aorta (AAA) endovascularly repaired by a stent graft (SG) system using the chimney technique. Computational fluid dynamics (CFD) is employed to study models of a healthy aorta versus an aorta post 'chimney' endovascular aneurysm repair (ChEVAR) using chimney stent grafts (CSG) inserted into each renal artery in parallel to the aortic SG. Results demonstrate that the presence of the CSGs results in stagnation regions and wall shear stress (WSS) modifications, yet the flow regime remains laminar. Thus, indicating the spatially contained effects of the ChEVAR technique and further supporting its merit.

Keywords: ChEVAR, abdominal aortic aneurysm (AAA), chimney stent grafts (CSG), computational fluid dynamics (CFD), hemodynamics, wall shear stresses (WSS)

1 Introduction

Aortic aneurysms are a main cause of death in the elderly population throughout the western world. The most common location for an aortic aneurysm formation is the infrarenal aorta (Guo *et al*, 2006). The traditional and most prevailing method of aneurysm repair is open surgery, whereby a large incision in the patient's abdomen allows access to the aneurysm site.

In recent years, more aneurysm repairs are performed endovascularly, excluding the aneurysm sac using stent grafts (SGs) inserted into the aneurysm site through the arterial system (minimally invasive). Typically, small incisions in the groin are created in order to deliver the SG system to the repair site using the femoral arteries as entry points. Following SG implantation, the aneurysm sac is sealed and blood subsequently flows through the newly created artificial conduit replacing the bulging part of the aorta.

A successful endovascular repair depends on the blood vessels and aneurysm morphologies. An aneurysm characterized by close proximity to a visceral artery ostium, is very challenging for endovascular repair. The requirement to properly seal the aneurysm sac while avoiding coverage of aortic branches by the SG can be very demanding. Innovative solutions for this type of problem include the fenestrated SG system. Fenestrated SGs are tailored to a specific patient morphology (Kandail *et al*, 2014).

In urgent cases, where the patient cannot wait several months for a custom SG system to be fabricated, an innovative solution is recently being employed using off-the-shelf SGs. This solution involves an endovascular surgical procedure called the 'chimney' technique whereby parallel to the main aortic SG that excludes the aneurysm sac, one or more tubular covered stents ('chimneys') are inserted into the visceral arteries. These covered stents facilitate proper blood flow to arteries that would otherwise be blocked by the main aortic SG due to their proximity to the aneurysm sac. A common case of aneurysm repair using the 'chimney' technique is the two renal arteries being highly adjacent to the aneurysm (Figure 1). Thus, requiring a chimney stent graft (CSG) in each renal artery to preserve blood flow to the kidneys.

In this study, a healthy abdominal aorta was evaluated in comparison with several configurations of post ChEVAR aorta having CSG inserted into each renal artery (Figures 1 and 2).

Computational fluid dynamics (CFD, ANSYS Fluent package) simulations of pulsatile blood flow during the cardiac cycle were employed. An idealized anatomy of the abdominal aorta was modeled based on averages of measurements taken from cadaver specimens and patient angiograms (Moore *et al*, 1992).

The effect of CSGs on abdominal aortic blood flow and wall shear stresses (WSS) was analyzed by examining blood flow patterns and regimes.

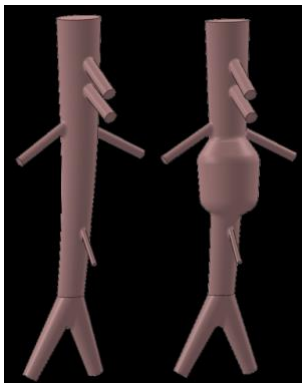


Figure 1: Left, healthy abdominal aorta model. Right, aneurysmatic aorta (aneurysm is adjacent to the renal branches).

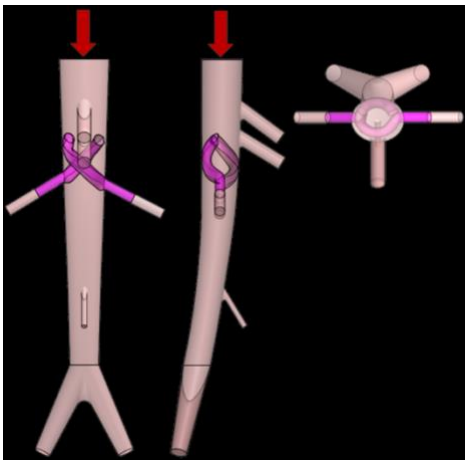


Figure 2: 3D model of the post ChEVAR abdominal aorta for analysis (aneurysm replaced by SG). Left to right: Front, side and top views, respectively. Red arrows: direction of blood flow.

2 Methodology

2.1 Governing Equations

The governing equations for blood flow in the abdominal aorta are the Navier Stokes momentum equations and the continuity equation for an incompressible fluid:

$$\rho \partial V/\partial t + \rho(V \cdot \nabla)V - \mu \Delta V + \nabla p = 0 \qquad (1)$$

$$\nabla \cdot V = 0 \qquad (2)$$

where $\underset{\sim}{v}$, ρ, μ & p are the fluid velocity, density, dynamic viscosity the pressure field experienced by the fluid, respectively. Blood is not a Newtonian fluid – viscosity depends on the strain rate according to the Carreau model for shear thinning fluids:

$$\mu(\dot{\gamma}) = \mu_\infty + (\mu_0 - \mu_\infty)(1 + \lambda^2 \dot{\gamma}^2)^{0.5(n-1)} \qquad (3)$$

where $\dot{\gamma}$ is the scalar flow shear rate and μ_∞ & μ_0 are the viscosities for an infinitely large and zero strain rate, respectively. λ and n are fluid specific time constant and power behavior index. Blood density is assumed 1045 kg/m^3 (Ene-Iordache *et al*, 2001).

2.2 Anatomical Model

The geometric 3D model employed for analyzing the idealized healthy abdominal aorta is presented in Figure 1. The model is based on angiograms and pressurized cadaver specimens measurements (Moore *et al*, 1992). The model incorporates the elliptical cross section, the tapering nature of the abdominal aorta, the arterial branches and the slight curvature towards the posterior wall. The model of the abdominal aorta post ChEVAR is based on the healthy model. Modifications were made in the model in order to account for the CSGs. The bulging part of the abdominal aorta is assumed completely replaced by the aortic SG, and thus is not a part of the numerical domain. The CSGs are modeled as long fabric-covered stents having a free diameter of 7 mm and a wall thickness of 0.1 mm, in compliance with suitable endograft dimensions often utilized in chimney repairs. The CSG models incorporate their helical-like nature (Coscas *et al*, 2011). A slightly flattened CSG region spanning from the orifice of the renal artery to the final contact region between the chimney and the aortic SG morphing the CSG cross section from a circle to an ellipse was also incorporated (de Bruin *et al*, 2013).

The CSGs protrude upstream into the aorta 10 mm above the main SG to avoid blockage of blood flow into the renal arteries.

2.3 Numerical Model

Blood flow behavior in the abdominal aorta during the cardiac cycle is considered to be predominantly laminar (Morris *et al*, 2004). Thus, a laminar CFD solver is employed. Literature demonstrates flow parameters e.g. WSS differ by as much as 30% between distensible and rigid blood vessel models (Shipkowitz *et al*, 1998). However, overall flow dynamics remain similar (Friedman *et al*, 1992). Thus, rigid wall approximation is sufficient for a comparative study.

No slip/penetration boundary conditions are applied at the walls. The inlet boundary condition employed is a pulsatile velocity function adapted from the literature - Figure 3 (Taylor *et al*, 1998). This waveform is decomposed into a Fourier series and modified to comply with the average velocity (flow rate). A parabolic profile distributed over the elliptical inlet is assumed (Shipkowitz *et al*, 1998). The domain has seven outlets with a constant flow ratio between them during the cardiac cycle (Moore *et al*, 1992). ANSYS Fluent CFD package (second order approximations) is used for the analysis.

2.4 Numerical Discretization

The model of the post ChEVAR abdominal aorta was meshed using 1.1 million polyhedral cells with 4 million nodes. The cycle time was discretized into 800 time steps. The scaled residuals value used was $5 \cdot 10^{-6}$. The numerical parameters used for the healthy aorta model were similar.

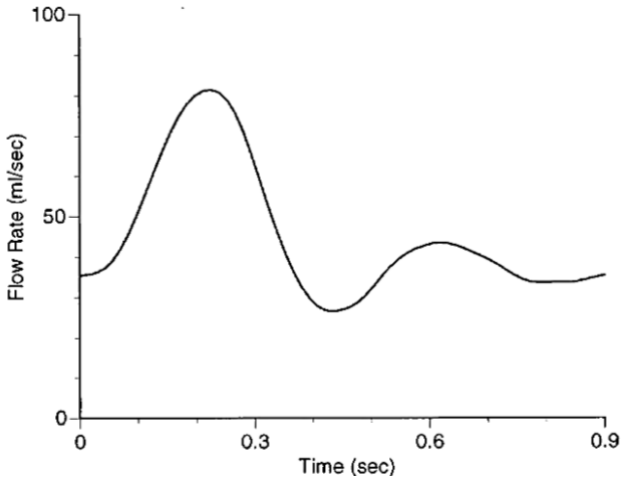


Figure 3: Waveform of the inlet flow rate [10].

3 Results

3.1 Validation

WSS for the healthy aorta (supra-celiac height) were compared with values measured in an experimental study (Moore *et al*, 1994).

WSSs were derived by extracting the temporal (during an entire cardiac cycle) minimum, maximum and average WSS values for each element of the supra-celiac ring of Figure 4. A spatial average of each parameter along the ring circumference was employed. Pulse WSS is defined as the difference between the maximum and minimum WSS for each element spatially averaged along the ring.

Results of this comparison are listed in Table 1 (Y/axial component). The relative errors indicate numerical results are in reasonable agreement with experimental data.

3.2 Flow Patterns

Stagnant regions are formed in the post ChEVAR aorta downstream near the CSGs. These regions persist throughout the cardiac cycle (Figures 5 and 6). There are no stagnant regions in the healthy model (Figures 7 and 8).

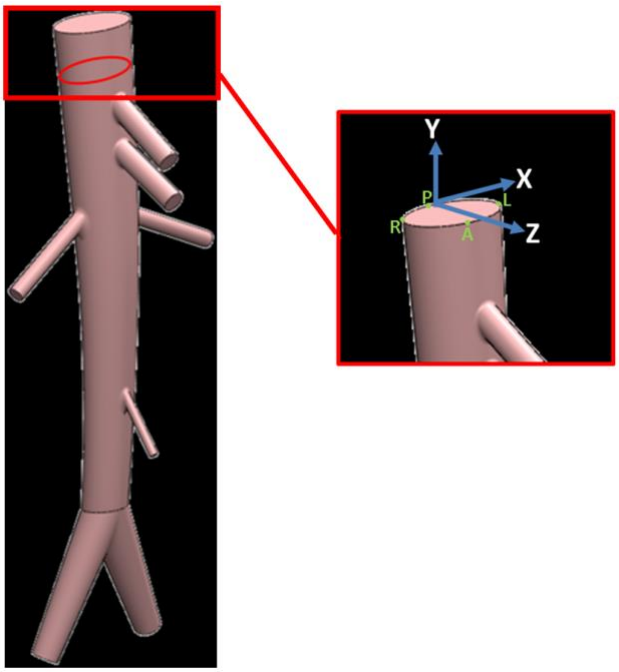


Figure 4: Left: WSS supraceliac comparison surface (ring in red). Right: Coordinate system and sectors of a horizontal ring of the abdominal aorta wall. A - anterior sector, P - posterior sector, R - right sector, L - left sector.

Table 1. Numerical validation results for wss values (y/axial component) in the supra-celiac region.

	Minimum WSS [Pa]	Maximum WSS [Pa]	Average WSS [Pa]	Pulse WSS [Pa]
Numerical Model	-0.48	0.99	0.19	1.47
Experiment	-0.45	0.87	0.13	1.32
Relative Error [%]	6	14	44	12

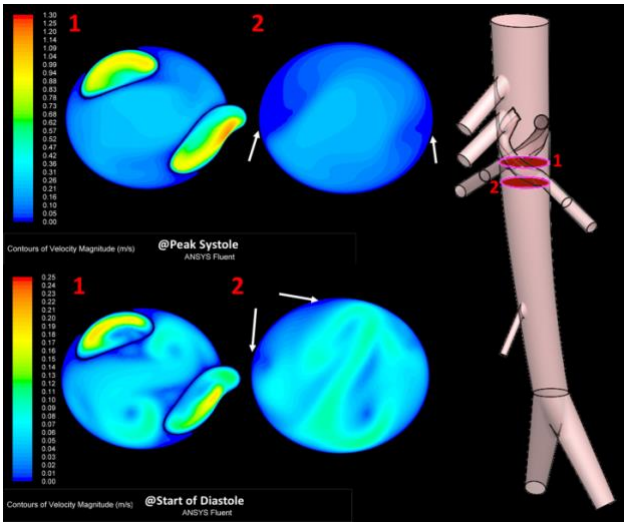


Figure 5: Velocity contours at peak systole and start of diastole at two distances below the CSGs (marked in red on the right). Arrows: stagnant regions.

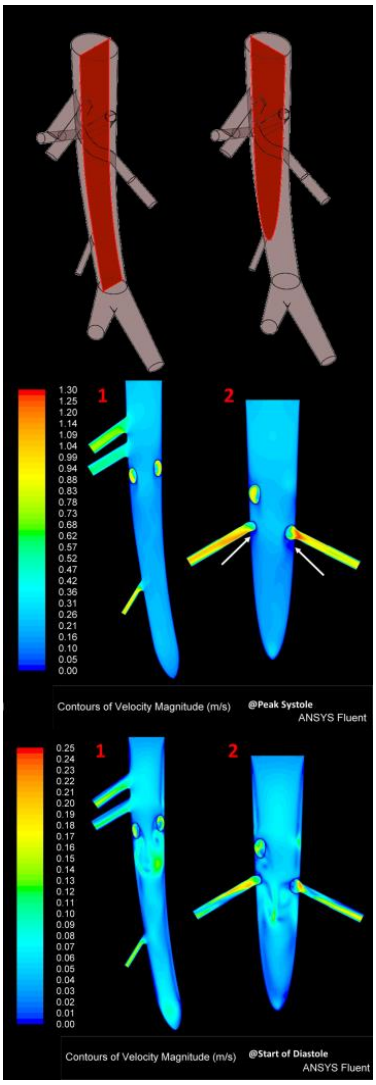


Figure 6: Post ChEVAR abdominal aorta. Top: planes of velocity contours are marked in red. Middle: Velocity contours at peak systole. Bottom: Velocity contours at diastole beginning.

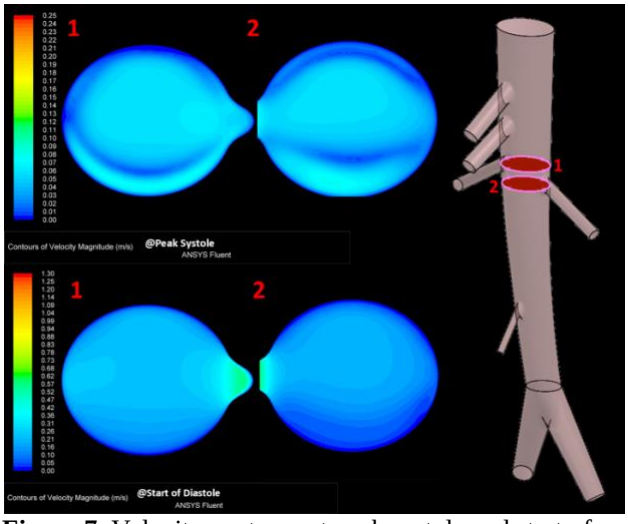


Figure 7: Velocity contours at peak systole and start of diastole at the same distances from the inlet as in Figure 5 (marked in red on the right) for a healthy aorta.

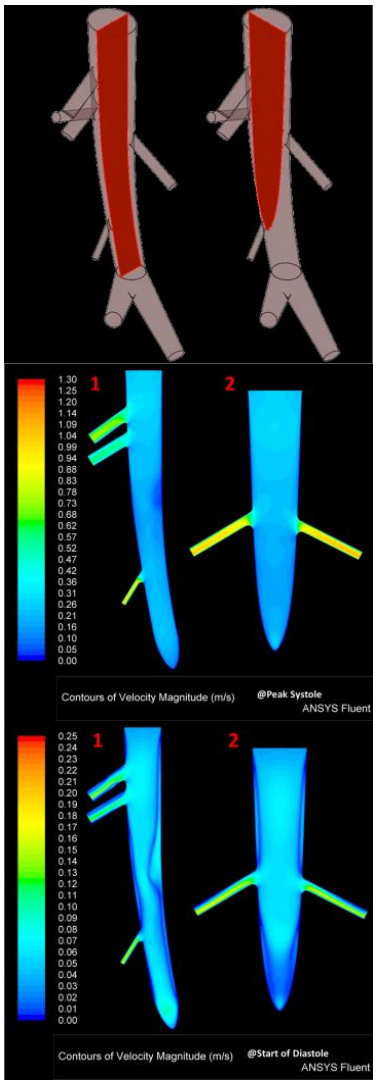


Figure 8: Healthy aorta. Top: planes of velocity contours are marked in red. Middle: Velocity contours at peak systole. Bottom: Velocity contours at diastole beginning.

3.3 Flow Regime

Y (axial) component of the WSSs for the aorta post ChEVAR at various positions and distances from the inlet (according to Figures 9 and 10) are plotted in Figure 11 through Figure 13. Y component of the velocity along the centerline is plotted in Figure 14. The WSSs and the velocity follow the inlet velocity waveform. There are no high frequency components present. If the inlet blood flow waveform is free of high frequency components yet points inside the control volume present velocity/WSS waveforms with high frequency noise then the flow exhibits transitional regime behavior. If the waveforms are free of high frequency components/noise, it indicates a laminar flow regime (Bozzetto *et al*, 2015). This implies that the flow in the post ChEVAR abdominal aorta is free of transitional behavior and is indeed laminar.

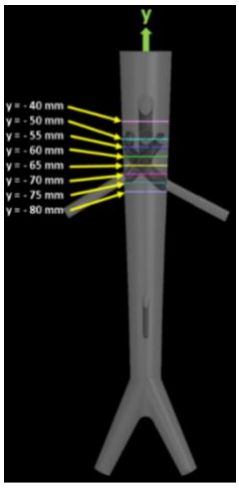


Figure 9: Section planes and their distances from the inlet.

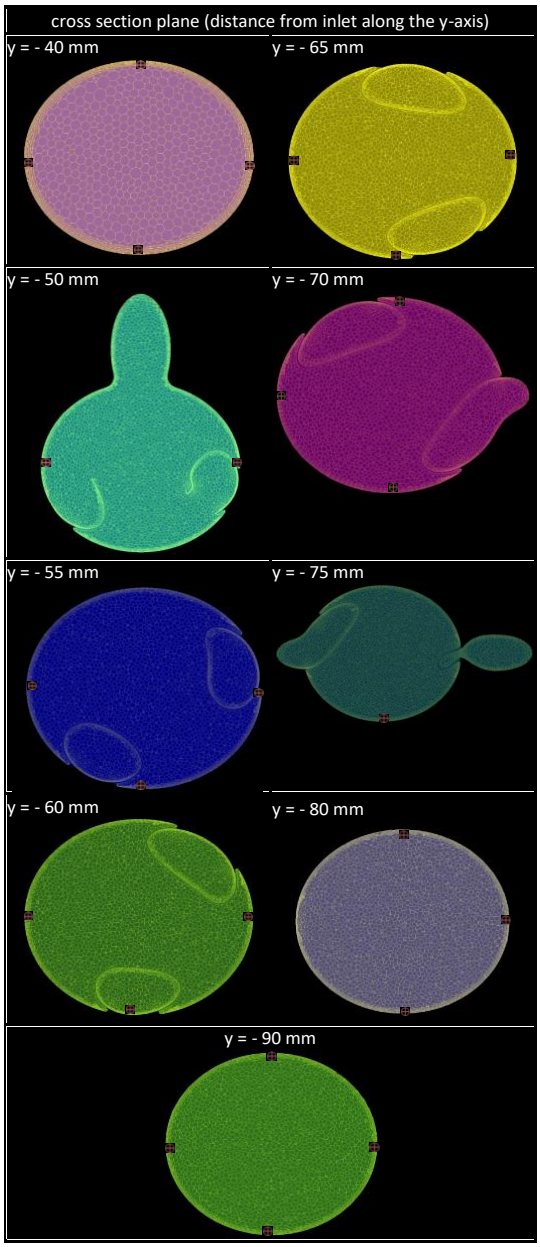


Figure 10: ⊕ Points evaluated for WSSs at different horizontal planes.

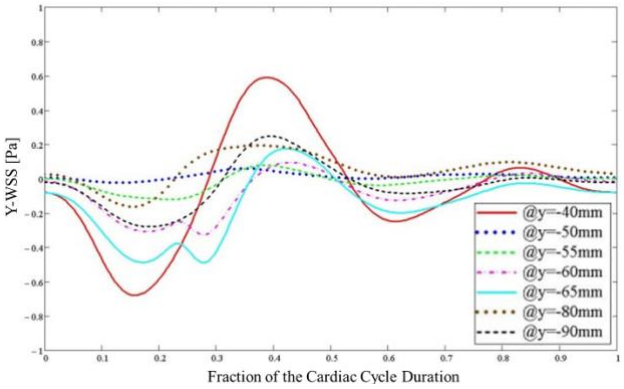


Figure 11: Y/axial component of WSS along the right side of the post ChEVAR aorta during the cardiac cycle.

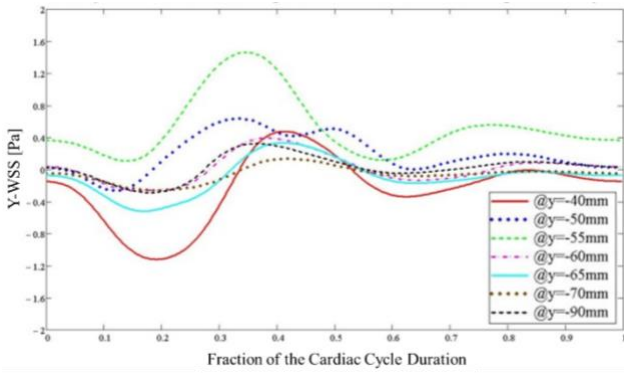


Figure 12: Y/axial component of WSS along the left side of the post ChEVAR aorta during the cardiac cycle.

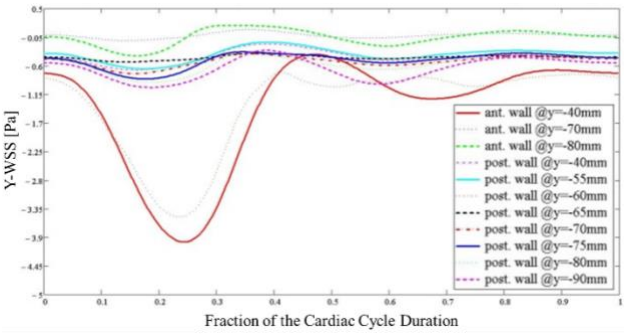


Figure 13: Y/axial component of WSS along the anterior and posterior of the post ChEVAR aorta during the cardiac cycle.

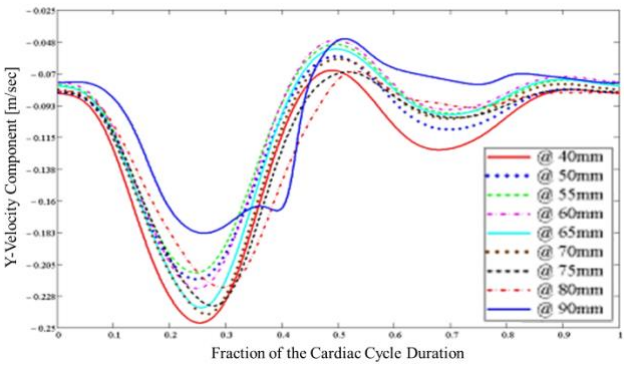


Figure 14: Y/axial component of velocity along the centerline of the post ChEVAR aorta.

4 Discussion and Conclusions

Our results suggest that CSGs presence in the abdominal aorta introduces variations in blood flow patterns and the formation of stagnant regions downstream from the CSGs throughout the cardiac cycle, potentially contributing to thrombosis (Ku *et al*, 1997). However, as can be deduced from the smooth and non-disturbed nature of the curves portrayed in Figure 11 through Figure 14 and in accordance with a previous study, the CSGs do not cause the flow regime to become turbulent or transitional (Bozzetto *et al*, 2015). This indicates limited flow field changes due to CSGs, thus further supporting the predictability of the flow in the abdominal aorta following an implantation of two renal stent grafts. These findings reconcile with data indicating a relatively high success rate in ChEVAR procedures performed in recent years, evident both in short and long-term patient follow ups (Zhang *et al*, 2015).

References

M. Bozzetto, B. Ene-Iordache, and A. Remuzzi. Transitional Flow in the Venous Side of Patient-Specific Arteriovenous Fistulae for Hemodialysis. *Ann. Biomed. Eng*., pp. 1–14, 2015. doi: 10.1007/s10439-015-1525-y.

R. Coscas, H. Kobeiter, P. Desgranges, and J.-P. Becquemin. Technical aspects, current indications, and results of chimney grafts for juxtarenal aortic aneurysms. *J. Vasc. Surg.*, 53(6):1520–7, 2011. doi.org/10.1016/j.jvs.2011.01.067.

J. L. de Bruin, K. K. Yeung, W. W. Niepoth, R. J. Lely, Q. Cheung, A. de Vries, and J. D. Blankensteijn. Geometric study of various chimney graft configurations in an in vitro juxtarenal aneurysm model. *J. Endovasc. Ther*., 20(2): 184–90, 2013. doi.org/10.1583/1545-1550-20.2.184.

B. Ene-Iordache, L. Mosconi, G. Remuzzi, and A. Remuzzi. Computational fluid dynamics of a vascular access case for hemodialysis. *J. Biomech. Eng*., 123(3): 284–292, 2001. doi: 10.1115/1.1372702.

M. H. Friedman, C. B. Bargeron, D. D. Duncan, G. M. Hutchins and F. F. Mark. Effects of Arterial Compliance and Non-Newtonian Rheology on Correlations Between Intimal Thickness and Wall Shear. *J. Biomech. Eng*., 114(3): 317, 1992. doi:10.1115/1.2891389.

D. N. Ku. Blood Flow in Arteries, Annu. Rev. *Fluid Mech*., 29: 399–434, 1997. doi.org/10.1146/annurev.fluid.29.1.399.

D.-C. Guo, C. L. Papke, R. He, and D. M. Milewicz. Pathogenesis of thoracic and abdominal aortic aneurysms. *Ann. N. Y. Acad. Sci*., 1085: 339–52, 2006. doi: 10.1196/annals.1383.013.

H. Kandail, M. Hamady, and X. Y. Xu. Patient-specific analysis of displacement forces acting on fenestrated stent grafts for endovascular aneurysm repair. *J. Biomech*., 47(14): 3546–3554, 2014.

J. E. Moore, D. N. Ku, C. K. Zarins and S. Glagov. Pulsatile flow visualization in the abdominal aorta under differing physiologic conditions: implications for increased susceptibility to atherosclerosis. *J. Biomech. Eng*., 114: 391–397, 1992. doi:10.1115/1.2891400.

J. E. Moore, S. Glagov, and D. N. Ku. Fluid wall shear stress measurements in a model of the human abdominal aorta: oscillatory behavior and relationship to atherosclerosis. *Atherosclerosis*, 9150: 225–240, 1994. doi.org/10.1016/0021-9150(94)90207-0.

L. Morris, P. Delassus, M. Walsh, and T. McGloughlin. A mathematical model to predict the in vivo pulsatile drag forces acting on bifurcated stent grafts used in endovascular treatment of abdominal aortic aneurysms (AAA). *J. Biomech*., 37(7): 1087–95, 2004. doi.org/10.1016/j.jbiomech.2003.11.014.

T. Shipkowitz, V. G. J. Rodgers, L. J. Frazin, and K. B. Chandran. Numerical study on the effect of steady axial flow development in the human aorta on local shear stresses in abdominal aortic branches. *J. Biomech*., 31: 995–1007, 1998.

C. A. Taylor, T. J. R. Hughes, and C. K. Zarins. Finite Element Modeling of Three-Dimensional Pulsatile Flow in the Abdominal Aorta: Relevance to Atherosclerosis. *Ann. Biomed. Eng*., 26: 975–987, 1998.

Y. Li, T. Zhang, W. Guo, C. Duan, R. Wei, Y. Ge, X. Jia, and X. Liu. Endovascular chimney technique for juxtarenal abdominal aortic aneurysm: a systematic review using pooled analysis and meta-analysis. *Ann. Vasc. Surg*., 29(6): 1141–50, 2015. doi.org/10.1016/j.avsg.2015.02.015.

Loadbalancing on Parallel Heterogeneous Architectures: Spin-image Algorithm on CPU and MIC

Ahmed Eleliemy[1] Mahmoud Fayze[2] Rashid Mehmood[3] Iyad Katib[3] Naif Aljohani[3]

[1]HPC Group, University of Basel, Basel, Switzerland, ahmed.eleliemy@unibas.ch
[2]Fujitsu & Computer Science, Ain-Shams University, Cairo, Egypt, Mahmoud.Fayez@ts.fujitsu.com
[3]High Performance Computing Center, King AbdulAziz University, Jeddah, Saudi Arabia, {rmehmood, iakatib, nraljohani}@kau.edu.sa

Abstract

Loadbalancing of computational tasks over heterogeneous architectures is an area of paramount importance due to the growing heterogeneity of HPC platforms and the higher performance and energy efficiency they could offer. This paper aims to address this challenge for a heterogeneous platform comprising Intel Xeon multi-core processors and Intel Xeon Phi accelerators (MIC) using an empirical approach. The proposed approach is investigated through a case study of the spin-image algorithm, selected due to its computationally intensive nature and a wide range of applications including 3D database retrieval systems and object recognition. The contributions of this paper are threefold. Firstly, we introduce a parallel spin-image algorithm (PSIA) that achieves a speedup of 19.8 on 24 CPU cores. Secondly, we provide results for a hybrid implementation of PSIA for a heterogeneous platform comprising CPU and MIC: to the best of our knowledge, this is the first such heterogeneous implementation of the spin-image algorithm. Thirdly, we use a range of 3D objects to empirically find a strategy to loadbalance computations between the MIC and CPU cores, achieving speedups of up to 32.4 over the sequential version. The LIRIS 3D mesh watermarking dataset is used to investigate performance analysis and optimization.

Keywords: heterogeneous architectures, MIC, spin-image algorithm, loadbalancing, performance analysis

1 Introduction

High performance computing (HPC) systems rely on concurrent, parallel and distributed computing technologies and resources to provide much larger memories and computational power than is possible with a general-purpose computer. HPC systems have traditionally been used to solve large problems arising from engineering and science. They are now being increasingly utilized in many other areas including business, economy and social sciences. Early HPC systems have mainly been homogeneous. However, the heterogeneity of modern computing systems is on the rise due to the increasing demands for higher performance and energy efficiency. Loadbalancing of computational tasks on homogeneous platforms is generally considered a difficult problem; it is an even bigger challenge when it comes to high performance heterogeneous systems.

Two most common options for accelerator units in heterogeneous platforms are GPGPUs (General-Purpose Computation on Graphics Processing Unit) and MICs (Intel Many Integrated Core Architecture). GPUs comprise thousands of cores and could offer very high memory bandwidth and computation throughput. In the last decade, a lot of research work has been done to speed up different algorithms using such architectures (Bautista Gomez et al., 2014). Therefore, GPUs have become one of the main accelerators in many HPC facilities. Although they are low-power, low-cost and massive parallel execution units but unfortunately, they have their own limitations (Shukla and Bhuyan, 2013). They are not compatible with existing x86 C, C++, and FORTRAN source codes. To use these architectures, we have to rebuild different codes from scratch and this could inhibit HPC users.

The MIC architecture was introduced by Intel in 2012 (Rahman, 2013). It is low-power, low-cost and massive parallel execution unit Like GPGPUs. However, it is a massively parallel architecture that is compatible with x86 applications. Programming models like Pthreads, MPI and OpenMP can be used without any code modifications (Utrera et al., 2015). Due to the programming simplicity and compatibility of the MIC architecture, it is the main accelerator unit in many modern HPC facilities (IntelPR). According to the Top500 list (http://www.top500.org/lists/) of most powerful supercomputers in the world, many fastest supercomputers are powered by Intel Xeon Phi coprocessors and Intel Xeon processors. There is a high potential of using MIC coprocessors beside Intel Xeon processors (Faheem and Konig-Ries, 2014). Therefore, we have chosen MICs alongside CPUs in this study of a heterogeneous platform.

This paper presents an empirical approach for achieving loadbalancing and optimum performance from a heterogeneous system comprising Intel Xeon multi-core processor and MIC. The approach is investigated using a case study of spin-image algorithm (Johnson, 1997). This algorithm is selected because it is being used in a wide range of important applications including 3D Database Retrieval Sys-

tems (Assfalg et al., 2004), Face Detection (Choi and Kim, 2013), Object Recognition (Johnson and Hebert, 1999), Object Categorization (Eleliemy et al., 2013), 3D map registration (Mei and He, 2013), and registration algorithm for LiDAR 3D point cloud models (He and Mei, 2015). The spin-image algorithm is well-known for its high computational complexity and is considered an essential bottleneck in many fields.

The paper makes three contributions. (1) A parallel spin-image algorithm (PSIA) has been introduced that achieves the speedup of 19.8 on 24 CPU cores. The process of parallelizing spin-image algorithm into a set of independent tasks which can be optimally scheduled between Intel Xeon processor and Intel MIC coprocessor is depicted and described. (2) results from a hybrid version of PSIA for a heterogeneous platform comprising CPU and MIC have been provided. (3) A strategy to empirically find an optimal loadbalancing of computations between the MIC and CPU cores has been introduced using a range of 3D objects. Speedups of up to 32.4 over the sequential version have been reported. We have used a range of objects from LIRIS 3D mesh watermarking dataset for performance analysis and optimization in all our experiments. To the best of our knowledge, no hybrid implementation of the spin-image algorithm on CPUs and MICs has been reported in the literature. The study of loadbalancing as a methodology for spin- over CPU and MIC is also novel.

This paper is organized as follows. Section 2 describes the sequential spin-image algorithm. Section 3 provides the dependency analysis of the spin-image algorithm and the proposed parallel spin-images algorithm. Section 4 provides details of the heterogeneous platform and dataset we have used for experiments in this paper. Results from the experiments and their analysis are presented along with the details of the loadbalancing strategy. Finally, in Section 5, further research challenges for loadbalancing on heterogeneous platforms and a number of directions for future work are given.

2 Spin-Image Algorithm

Spin-image is an algorithm that converts a 3D shape into a set of 2D images as shape descriptors. It was introduced in (Johnson, 1997). The main concern regarding the use of the spin-image algorithm is its computational complexity, especially with the increase of depth cameras' resolution. According to the spin-image algorithm, a spin-image can only be generated at any point with known normal (oriented point). Generated spin-image can be viewed as a paper that rotates around point normal while other points touch and stick to it.

$$i = \frac{(W/2) - n \cdot (x - p)}{B} \tag{1}$$

$$j = \frac{\sqrt{||x-p||^2 - (n.(x-p))^2}}{B} \tag{2}$$

Equations 1 and 2 show how to calculate spin-image at an oriented point (p). This equation calculates two indices i and j, where the generated spin-image should be incremented by one. In fact, instead of incrementing the point at index i and j, four indices $[i, j]$, $[i, j+1]$, $[i+1, j]$ and $[i+1, j+1]$ are incremented with values $(1-a)(1-b)$, $(1-a)b$, $a(1-b)$ and ab, respectively. Equations 3 and 4 can be used to calculate the values of a and b. This process is called smoothness of spin-images.

$$a = a - i * binsize \tag{3}$$

$$b = \beta - j * binsize \tag{4}$$

The equations show that i and j are affected by two parameters; Image-width (W) and Bin-size (B). However, these parameters affect only the quality of generated spin-image and do not have impact on spin-image generation time. Moreover, smoothness process is related to the quality of the generated spin-image. Therefore, these parameters will be considered as constants and smoothness process will be ignored in this work. The below pseudo code shows the generation process spin-images .

```
1  procedure CalcSpinImages
2  W = ImageWidth
3  B = BinSize
4  Define SpinImages As List
5  for each p in Mesh
6    Define SpinImage[W * W]
7    n = normal of P
8    for each x in Mesh
9       i = ((W/2)-n×(x-p))/B
10      j= √(||x-p||²-(n.(x-p))²)/B
11      SpinImage[i,j]+=1
12   end for
13   SpinImages.add(SpinImage)
14 end for
15 return SpinImages
16 end procedure
```

3 The Proposed Parallel Spin-Images Algorithm

Figure 1 shows the spin-image generation process for a given 3D Mesh. This process contains different levels of calculations. Each level contains some of the dependent and/or independent tasks. For example, Let M be a certain 3D Mesh, S is the set of all M oriented points, P_1, P_2 and $P_n \in$ S. Spin-image calculation at P_1 is totally independent of P_2, P_3 and P_n. However, at P_1, we have to scan P_2, P_3 till P_n to find different indices i, j to increment spin-image at these indices. Such calculation cannot be parallelized directly because it may be required to increment the same spin-image cell $[i, j]$ simultaneously.

Figure 2 shows the proposed PSIA. In PSIA, we have two levels of parallelism. At the first level, we calculate spin-images at different points P_1, P_2 and P_n simultaneously. While at the second level, for each point pair (P_x,

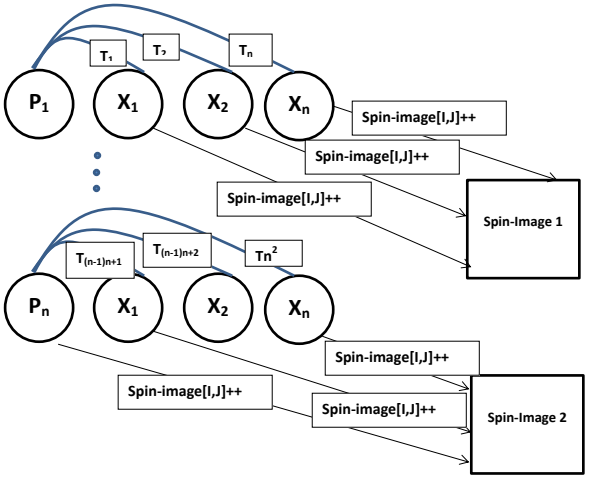


Figure 1. Generating Spin-images for a given 3D Mesh

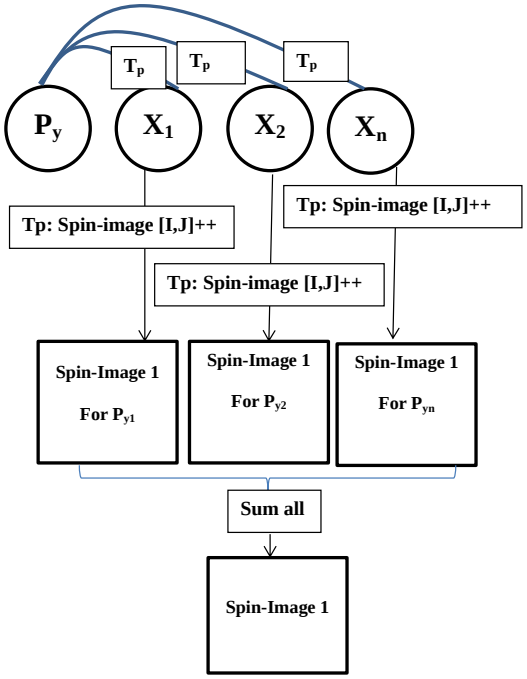


Figure 2. Proposed Parallel Spin-Image Algorithm

P_1) , (P_x, P_2) and (P_x, P_n), we have a temporary spin-image that we call *Partial Spin-image*. Finally, we add all partial images to get the spin-image at point P_x. Figure 2 shows the process for computing all spin-images concurrently using partial spin-images.

4 Results and Analysis

4.1 Platform Specification

The experiments have been carried out on the *Aziz* supercomputer. Aziz supercomputer is Fujitsu made and is able to deliver peak performance of 230 teraflops. It has a total of 11,904 cores in 496 nodes, where each node comprises dual socket Intel Xeon E5-2695v2 12-core processor running at 2.4GHz. 380 of these nodes contain 96 GB memory each, while the rest of the 112 nodes contain 256 GB each, making up a total of 66 TB memory in the system. The system also contains 2 NVidia Tesla K20 GPU equipped compute nodes with 48 cores and 2 Intel Phi 5110P co-processor equipped compute nodes with 48 cores. Aziz was ranked number 360 in the June 2015 Top500 competition, currently it is at number 491 (November 2015).

The platform (part of the Aziz supercomputer) we have used for the experiments consists of 2 Intel processors E5-2695v2 each has 12 Cores (2.4GHz), 96GB RAM, Intel Xeon Phi Coprocessor 5110P (1053GHz, 60Cores). The code is written in C and compiled for Linux (CentOS 6.4) using Intel parallel studio XE 2015 version 15.0.

4.2 Experimental Data

Objects from LIRIS 3D mesh watermarking dataset (Wang et al., 2010) have been used for the performance analysis and optimization. This dataset has been selected due to its dense objects. Table 1 shows the number of vertices for each object.

Table 1. 3D Objects in 3D Mesh Watermarking Dataset

Object Label	*Number of vertices*
Ramesses	826266
Horse	112642
Venus	100759
Rabbit	70658
Crank	50012
Dragon	50000
Hand	36619
Bunny	34835
Casting	5096
Cow	2904

4.3 Results

It can be seen from our discussions in Sections 2 and 3 that the execution time of spin-images generation process for any object is the key measure of the computational complexity. As mentioned in section 2, the complexity of such process is $O(n^2)$. Therefore, any increase in the number of object points leads to significant increase in the total time of spin-image generation process. In order to illustrate that fact, only percentage of 1% spin-images for each object has been generated. In figure3, the time to spin-images generation for *Bunny* and *Rabbit* is 0.73 and 2.881 seconds respectively, while their sizes are 34835 and 70658 points respectively. Simply, *Rabbit* is almost 2 times in size comparing with *Bunny*, but its spin-images generation takes almost 4 times as what it takes in the *Bunny* object. Also, it is the same for both *Horse* and *Bunny*, time to generate spin-images for *Horse* is 9 times as *Bunny*, while its

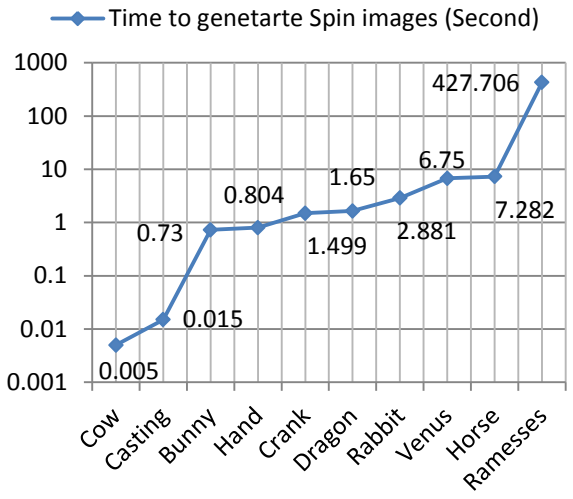


Figure 3. Sequential implementation of spin-image algorithm which number of generated spin-images equals to 1% of object size

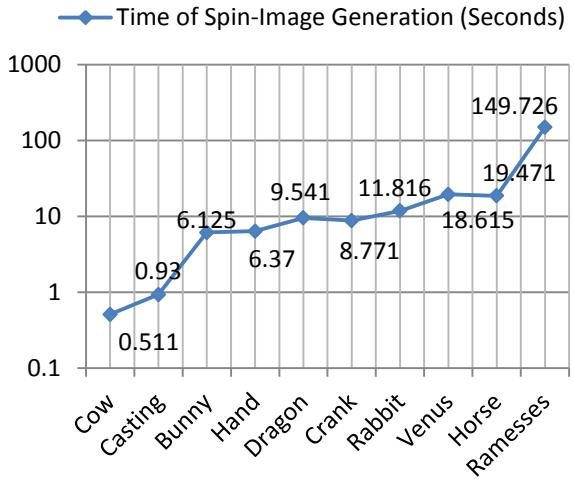


Figure 4. Sequential implementation of spin-image algorithm and the number of generated spin-images equals to 2904

is 3 times as *Bunny*. Some approaches from Object Recognition field as well as Object Categorization like (Hegazy, 2016) try to avoid such problem by avoiding spin-image generation at each point of a 3D object, it only generate spin-images at certain points which will not affect correct recognition or categorization rate. However, the problem remains specially for dense objects like LIRIS objects.

In Figure 4, the total number of generated spin-images is 2409 images per each object, this number represents the common maximum number of spin-images between all objects, because *Cow* Object is the smallest object and it contains 2409 vertices. Figure 4 shows that dense objects like *Ramesses*, *Horse*, and *Venus* consume at huge time such it can not be used in real-time systems. For example, *Horse*, *Venus*, and *Rabbit* objects take 19.471, 18.615, and 11.816 seconds respectively. Moreover *Ramesses* object takes 149.726 seconds.

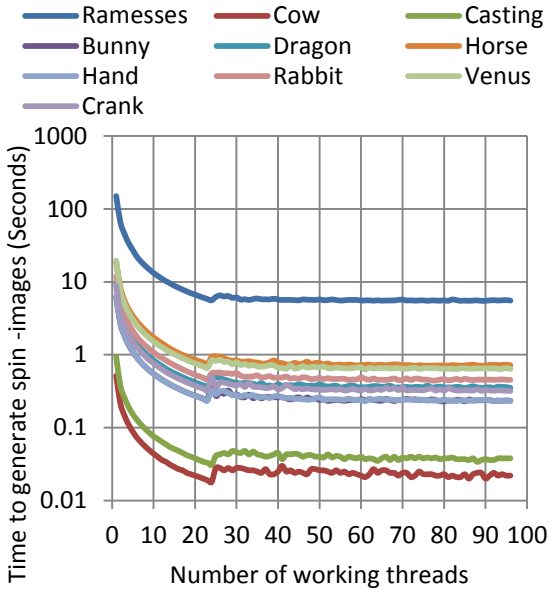


Figure 5. Performance of PSIA over 24-cores CPU and the number of generated spin-images is 2904

In fact, the *Ramesses* object is 284.5 times in size compared to the smallest object in the dataset. Therefore, it will always be a challenge to generate all of its spin-images, taking into consideration that all points of any LIRIS object are oriented points, which means that for *Ramesses* we need to generate 826266 spin-images.

Figure 5 shows the performance of our enhanced PSIA algorithm for multi-core CPUs. Note that As discussed in section 3, PSIA is a parallel version of the original spin-image algorithm where openMP threads are used to perform the parallel tasks. The figure shows that there is a reverse relation between the number of OpenMP threads and the total run-time. Simply, increasing the number of working OpenMP threads will decrease the total run-time. However, there is a turnover point on OpenMP threads axis at 24 thread, where this relation is not valid. Such relation violation is due to the physical characteristics of the used hardware, as mentioned before this experiment runs over 24-Cores CPU.

Another experiment has been conducted to investigate the scalability of the proposed parallel SIA algorithm using another shared memory architecture like Xeon phi (MIC technology). As mentioned before, we have an Intel Xeon Phi card that consists of 60 cores. There is two important feature in such hardware; First each core can run 4 concurrent threads. Second, there is no need to change any code, it is compatible with x86 processors. In other words, same code can be recompiled. Figure 6 shows same performance which means scalability of the algorithm, but it also shows the turnover point changed to be around 230 core which is logically due to the physical characteristics of the hardware platform.

Actually due to the type of spin-image calculations which all are floating point operations, results of running

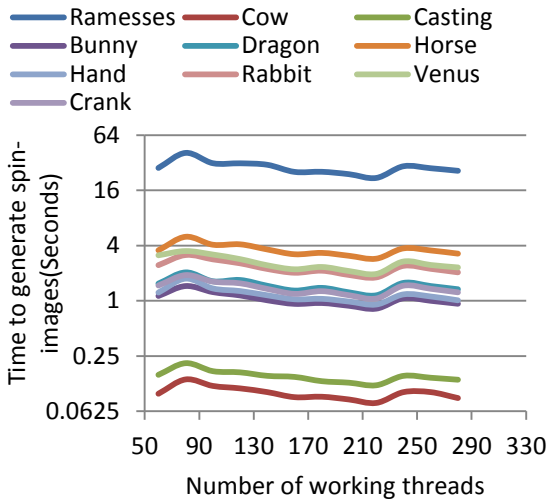


Figure 6. Performance of PSIA over 60-cores MIC accelerator and the number of generated spin-images is 2904

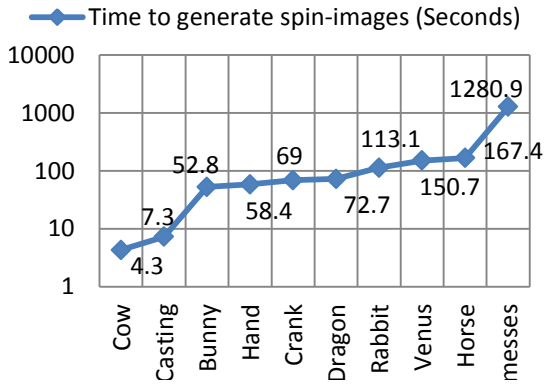


Figure 7. Sequential implementation of spin-image algorithm runs over MIC accelerator and the number of generated spin-images is 2904

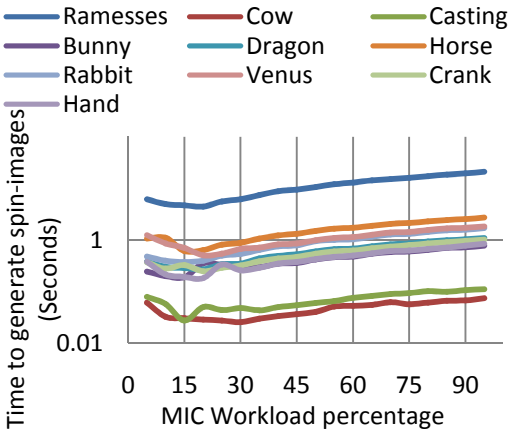


Figure 8. Performance for the hybrid implementation of PSIA at different workload distributions between CPU and MIC-cores

Table 2. The values for MIC Workload Ratios for the given objects

Object Name	R_1	R_2	*MIC Workload %*
Ramesses	0.232666667	0.095204348	29.04
Cow	0.00075	0.000334783	30.86
Casting	0.001291667	0.000521739	28.77
Bunny	0.017833333	0.003573913	16.69
Dragon	0.019291667	0.004995652	20.56
Horse	0.038833333	0.012530435	24.39
Hand	0.0135	0.003965217	22.70
Rabbit	0.02325	0.007830435	25.19
Venus	0.034875	0.00853913	19.67
Crank	0.020583333	0.004608696	18.29

PSIA over MIC is not promising as running over CPU, however, the total time to generate 2904 spin-images is reduced comparing to sequential spin-image algorithm. In order to complete the image, the original sequential spin-image has been compile to MIC architectures. Figure 7 shows that how the original sequential spin-image algorithm behave for MIC. Therefore, it expected that PSIA over MIC is not promising as running over CPU.

It can be observed from the results so far that the PSIA performance over CPU is better than its performance over MIC. However, using both CPU and MIC together should give better results than using any of them individually. The remaining question is how to divide the workload between CPU and MIC. According to (Faheem and Konig-Ries, 2014), because all PSIA tasks are identical with almost no dependency, workload for MIC can be calculated as follows.

Let R_1 is the ratio between the number of generated spin-images and run-time over 24 CPU cores, and R_2 is the ratio between the number of generated spin-images and run-time over 230 MIC cores. Formally, R_x = (the number of generated spin-images / total run time) / the number of working threads, where x represent the available architectures ($CPU = 1$ and $MIC = 2$) The MIC Workload Ratio W of MIC to the total workload could be calculated as $R_2/(R_1 + R_2)$. The number of generated spin-images is same for both R_1 and R_2, and therefore, it could be excluded from the workload Ratios (W). The values for MIC Workload Ratios (as well as R_1 and R_2) for various objects are given in Table 2. The table shows that the lowest MIC Workload Ratio (Column 4) is for the Bunny object (16.69%), while the highest is for the Cow object (30.86%). Consequently, based on the results given in Table 2, we can conclude that the maximum performance can be obtained by allocating around 23% (average of the values in Column 4) of the total workload to MIC and the remaining part to CPU. This is not the exact optimum value, rather a value around which optimum performance can be found.

To investigate the workload distribution between MIC and CPU further, we performed another set of experi-

ments. Figure 8 shows the results for a range of workload distributions between MIC and CPU. The first result (the leftmost) is for the case where MIC is given 5% of the workload, while CPU gets 95% of the workload. The MIC workload is increased in steps of 5% until it reaches 95%, where CPU gets a 5% of the total load. Based on the numerical values of the results depicted in the figure, the optimum MIC Workload Ratio is dependent on the object, and falls between 10% to 30%. Although further analysis of such behavior is required, our preliminary explanation is as follows. According to Equations (1), (2), (3), and (4), the values of i, j require 4 memory accesses — $[i, j], [i, j+1], [i+1, j]$ and $[i+1, j+1]$ — in order to increment the value of the spin-image. These memory accesses are avoided when the values j, j are outside the spin-image boundary. However, there are no guarantees that the MIC in the assigned workload may require, or may not require, these 4 memory accesses. This memory access behavior will be examined further in our future work.

5 Conclusions and future work

Improving loadbalancing of computational tasks over heterogeneous architectures is an area of paramount importance. This paper aimed at addressing the loadbalancing problem for a heterogeneous platform comprising CPUs and MICs. The approach proposed in this paper was investigated through a case study of the spin-image algorithm, selected due to its computationally intensive nature and a wide range of applications including 3D database retrieval systems and object recognition. The paper made three contributions. A parallel spin-image algorithm (PSIA) was introduced and its implementation achieved the speedup of 19.8 on 24 CPU cores. Results from a hybrid implementation of PSIA were presented. We empirically found a strategy to loadbalance computations between the MIC and CPU cores, achieving speedups of up to 32.4. Objects from LIRIS 3D mesh watermarking dataset were used to provide performance analysis and optimization.

The proposed PSIA on the heterogeneous platform can replace the original spin-image for many different applications as mentioned in Section 1. Also, the proposed PSIA is scalable. It can run over a different number of cores equal to N, such that N is less than or equal to the number of generated spin-images. The use of the PSIA algorithm is recommended for the case where the objects are dense. Moreover, the proposed PSIA hybrid implementation allows real-time generation of spin-images.

The results and analysis of our approach for loadbalancing on heterogeneous platforms show great promise. However, there are some concerns regarding the proposed PSIA that need further investigation. For example, PSIA is based on creating partial spin-images, which means that higher memory resources are required. Further investigation in needed to find out the memory profile of the proposed PSIA algorithm compared to the original version? The loadbalancing strategy needs to be investigated further with a wider range and size of objects. Analytical or heuristic formulations need to be devised. We need to investigate various cases where, for instance, one of the computing resources does not have sufficient memory to take its workload: is it better to distribute the workload based on the characteristics of the work or the amount of work? Finding answers to these questions forms our motivation for the future work. We also plan to add the energy efficiency dimensions to this work. We plan to look at optimizing the loadbalancing strategy against energy efficiency, memory profile and computational performance.

Acknowledgment

This work is supported by the King AbdulAziz University's High Performance Computing Center (`http://hpc.kau.edu.sa`). The spin-image computations described in this paper are performed on the *Aziz* supercomputer, part of the HPC Center.

References

J. Assfalg, G. D'Amico, A. Del Bimbo, and P. Pala. 3D content-based retrieval with spin images. In *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, volume 2, pages 771–774, June 2004. doi:10.1109/ICME.2004.1394314.

L. Bautista Gomez, F. Cappello, L. Carro, N. DeBardeleben, B. Fang, S. Gurumurthi, K. Pattabiraman, P. Rech, and M. Sonza Reorda. GPGPUs: How to combine high computational power with high reliability. In *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*, pages 1–9, March 2014. doi:10.7873/DATE.2014.354.

K. S. Choi and D. H. Kim. Angular-partitioned spin image descriptor for robust 3D facial landmark detection. *Electronics Letters*, 49(23):1454–1455, Nov 2013. ISSN 0013-5194. doi:10.1049/el.2013.1577.

A. Eleliemy, D. Hegazy, and W.S. Elkilani. MPI parallel implementation of 3D object categorization using spin-images. In *Computer Engineering Conference (ICENCO), 2013 9th International*, pages 25–31, Dec 2013. doi:10.1109/ICENCO.2013.6736471.

H. M. Faheem and B. Konig-Ries. A new scheduling strategy for solving the motif finding problem on heterogeneous architectures. *International Journal of Computer Applications*, 101(5), September 2014.

Y. He and Y. Mei. An efficient registration algorithm based on spin image for Lidar 3D point cloud models. *Neurocom-puting*, 151, Part 1:354 – 363, 2015. ISSN 0925-2312. doi:http://dx.doi.org/10.1016/j.neucom.2014.09.029.

D. Hegazy. Symmetric multi-processing 3d object categorization model using a spin-point curvature selection strategy. *Egyptian Computer Science (ECS) Journal*, 40(1):73–83, January 2016.

A. Johnson. *Spin-Images: A Representation for 3-D Surface Matching*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, August 1997.

A.E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(5):433– 449, May 1999. ISSN 0162-8828. doi:10.1109/34.765655.

Y. Mei and Y. He. A new spin-image based 3D map registration algorithm using low-dimensional feature space. In *Information and Automation (ICIA), 2013 IEEE International Conference on*, pages 545–551, Aug 2013. doi:10.1109/ICInfA.2013.6720358.

Intel Newsroom. *Intel Delivers New Architecture for Discovery with Intel Xeon Phi Coprocessors*. Available via https://www.bayesfusion.com/ [accessed November 8, 2016].

R. Rahman. *Intel® Xeon Phi™ Coprocessor Architecture and Tools. The Guide for Application Developers*. Apress Media, ISBN13: 978-1-4302-5926-8, 2013.

S.K. Shukla and L.N. Bhuyan. A hybrid shared memory heterogeneous execution platform for PCIe-based GPDGUs. In *High Performance Computing (HiPC), 2013 20th International Conference on*, pages 343–352, Dec 2013. doi:10.1109/HiPC.2013.6799140.

G. Utrera, M. Gil, and X. Martorell. In search of the best MPI-OpenMP distribution for optimum Intel-mic cluster performance. In *High Performance Computing Simulation (HPCS), 2015 International Conference on*, pages 429–435, July 2015. doi:10.1109/HPCSim.2015.7237072.

K. Wang, G. Lavoué, F. Denis, A. Baskurt, and X. He. A benchmark for 3D mesh watermarking. In *Proc. of the IEEE International Conference on Shape Modeling and Applications*, pages 231–235, 2010.

CFD Approaches for Modeling Gas-Solids Multiphase Flows – A Review

W.K. Hiromi Ariyaratne [1] E.V.P.J. Manjula [1] Chandana Ratnayake [2] Morten C. Melaaen [1]

[1] Faculty of Technology, Natural Sciences and Maritime Sciences, University College of Southeast Norway, Post box 235, N-3603 Kongsberg, Norway,
hiromi.ariyaratne, jagath.m.edirisinghe, morten.c.melaae @usn.no

[2] Department of POSTEC, Tel-Tek, Kjølnes ring 30, N-3918, Porsgrunn, Norway, chandana.ratnayake@tel-tek.no

Abstract

This review study focuses on the application of Computational Fluid Dynamics (CFD) in the investigation of gas-solids multiphase flow systems. The applicability and limitations of conventional models and recent developments of existing multiphase models for the prediction of gas-solids flows are thoroughly overviewed. Use of conventional Eulerian-Eulerian model for granular flows and Lagrangian approach incorporated with Discrete Element Method (CFD-DEM) are quite well proven, however some limitations restrict the use of these models in wide range of applications. Therefore, some new models have been introduced to model gas-solids flows, as example Dense Discrete Phase Model incorporated with Kinetic Theory of Granular Flow (DDPM-KTGF), Dense Discrete Phase Model incorporated with Discrete Element Method (DDPM-DEM) and Computational Particle Fluid Dynamics (CPFD) numerical scheme incorporated with the MultiPhase-Particle-In-Cell (MP-PIC) method. These models have been validated for certain applications under certain conditions, however, further validation of these models is still a necessity.

Keywords: models, CFD-DEM, DDPM-KTGF, DDPM-DEM, MP-PIC

1 Introduction

Applications involving gas-solids multiphase flows are very common in numerous industrial processes and also in various natural phenomena, such as sand storms and cosmic dusts (Li et al., 2012). Pneumatic conveying units, hoppers, solids separation units such as cyclones, bubbling and circulating fluidized beds used in gasification, carbon capture, etc. can be identified as some of the industrial process units involved in gas-solids flows. To optimize the design and operation of industrial processes and also to understand natural phenomena which involve gas-solids flows, a thorough understanding of gas-solids flows is needed.

Achievement of this understanding involves the development of experimental measurement techniques, experimentally verified multiphase flow equations and numerical simulation tools (Arastoopour, 2001). Significant effort has been devoted to improving numerical tools, such as Computational Fluid Dynamics (CFD) tool, to predict such complex flows. However, it has been identified that systems containing one or more particulate phases are the most complex and challenging in the field of multiphase flow modeling. To accurately predict the solids behavior, it is necessary to choose a numerical method capable of accounting not only particle-fluid interactions but also for particle-wall and particle-particle interactions in three dimensions and across any particle size distribution (Parker et al., 2013).

Different types of CFD models are available for the prediction of gas-solids flows. Each model has inherent merits and disadvantages. Therefore, a certain model can be appropriate over another depending on the factors prioritized by the user e.g. accuracy of the results, computational time, applicability in large-scale systems, etc. Moreover, the models are still far from perfect and the available models are undergone many improvements within the time. In this review paper, some modeling approaches available for the modeling of gas-solids flow systems are analyzed including their applications and limitations. First, an overview of the models is presented. Then, the two basic approaches and the different models available under basic approaches are discussed.

2 Basic CFD Approaches for Modeling of Gas-Solids Flows

A brief summary of the discussed approaches and models are presented in Figure 1. In dealing with modeling of gas-solids flows, the Eulerian-Eulerian and the Eulerian-Lagrangian methods are the frequently used approaches (Chen and Wang, 2014). In Eulerian-Eulerian approach, all the phases are treated as continuous phases while in Eulerian-

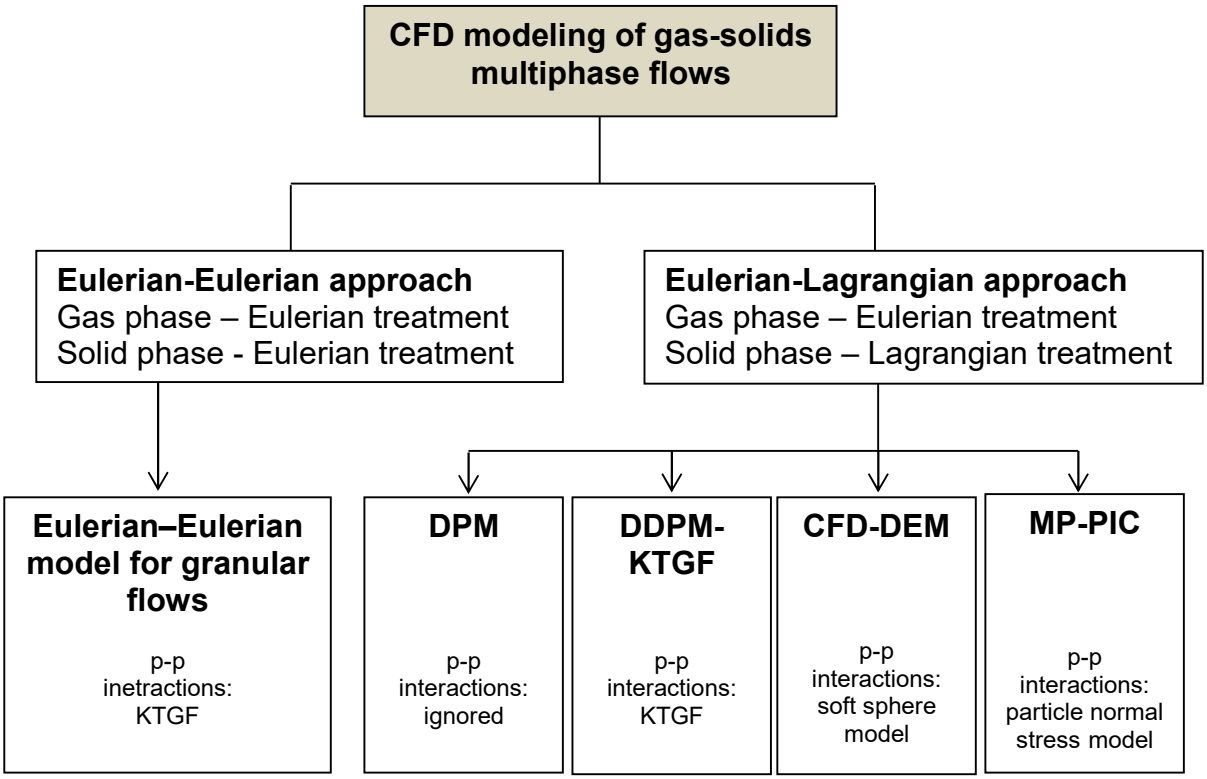


Figure 1. Summary of model approaches for gas-solids multiphase flow modelling.

Lagrangian approach, the fluid phase is treated as a continuous phase, but the solid phase is treated as a discrete phase. Eulerian-Eulerian model for granular flows is described under Eulerian-Eulerian approach and there are four main models under Eulerian-Lagrangian approach, namely, Lagrangian Discrete Phase Model (DPM), Dense Discrete Phase Model incorporated with Kinetic Theory of Granular Flow (DDPM-KTGF), CFD-Discrete Element Method (CFD-DEM) and Computational Particle Fluid Dynamics (CPFD) numerical scheme incorporated with the MultiPhase-Particle-In-Cell (MP-PIC) method.

In addition to the difference in the way of solid phase treatment, another basic difference of the models under both approaches is, the way of treating particle-particle (p-p) interactions. DPM neglects the p-p interactions, and other models consider the p-p interactions through different approaches such as kinetic theory of granular flow, particle normal stress model, soft sphere model, etc.

Much information about the model approaches are discussed in the following sections.

3 Eulerian-Eulerian approach

In the Eulerian-Eulerian approach, fluid and particles both are considered as continuous phases which are fully inter-penetrating (Zhang et al., 2012) i.e. solid phase is treated as a pseudo-fluid (Abbasi et al., 2013). Volume fractions of phases are assumed to be continuous functions of space and time. Since the volume of a phase cannot be occupied by the other phases, the sum of volume fractions is equal to one. This is the concept of the phasic volume fraction (Abbasi et al., 2011). The conservation equations of mass, momentum and energy for the phases are then obtained through an appropriate averaging process (typically ensemble-averaging) (Chen and Wang, 2014). The averaging procedure leads to many unclosed terms, which must be modeled (Snider et al., 2011). Constitutive relationships that are obtained from empirical information and/or kinetic theory are used for this purpose (Abbasi et al., 2013). Eulerian-Eulerian model for granular flows (Euler-granular model) is an example for Eulerian-Eulerian approach. As mentioned (Garg et al., 2012), commercially available codes like ANSYS Fluent, and open source codes like CFDlib, OpenFOAM® and MFiX are all capable of performing Eulerian-Eulerian simulations. Similar forms of governing equations are solved in all these codes and main difference can be found in closures for various sub models (such as solids stresses, interphase drag, etc.) and in numerical treatment.

The Eulerian-Eulerian approach normally requires less computational resources compared to Eulerian-Lagrangian approaches (Chen and Wang, 2014). And this approach is quite traditional and has played a very

important role in determining the fluid dynamic characteristics of gas–solids flow (Chen and Wang, 2014). Therefore, Eulerian-Eulerian approach has a wide application in gas–solids flows (Weber et al., 2013). However, this approach has major limitations in considering variations of particle properties, as example, wide particle size distribution, density diversification and sphericity consideration. Nevertheless, the particle size differences and/or density variations can affect the gas–solids flow behaviors such as solid segregation (Wang et al., 2014a), hence cannot be neglected in certain situations. In that case, many separate continuity and momentum equations are required to accurately represent the different particle types and sizes in this model (Andrews and O'Rourke, 1996). However, the computational cost of inclusion of many phases cannot then be overlooked, in fact it depends on the computational capacity available. Some researchers have quoted that the Euler-granular model cannot easily account for some characteristics of realistic particles such as shear stresses and inter-particle cohesive forces for Geldart A particles when treated as a pseudo fluid (Chen et al., 2013). Moreover, many researchers emphasize that incorporating of dissipation in the Kinetic Theory of Granular Flow (KTGF) model considering the effects of wall roughness is an important factor for the accurate prediction of results in Eulerian-Eulerian model for granular flows (Chen and Wang, 2014).

4 Eulerian-Lagrangian Approach

In Eulerian–Lagrangian approach, the fluid phase is still modeled with time-averaged Navier–Stokes equations and other conservation equations (Yin et al., 2014). The dispersed (solid) phase is treated by tracking a large number of particles through the calculated flow field (Abbasi et al., 2012). Each particle is affected in its trajectory by three-dimensional forces and Newtonian equations of motion are used for the calculations (Yin et al., 2014). Commercially available codes like ANSYS Fluent and Barracuda®, and open sources codes like MFiX-DEM, KIVA, OpenFOAM® are capable of performing Eulerian-Lagrangian simulations (Garg et al., 2012). The way of treating particle-particle interactions and the numerical method used to solve the equations are the main differences in different Eulerian-Lagrangian codes.

Compared to Eulerian-Eulerian approach, Eulerian-Lagrangian approach can provide analysis of flows with a wide range of particle types, sizes, shapes and velocities (Lu et al., 2014). However, if details of particle–particle and particle-wall collisions are explicitly tracked, the traditional Lagrangian models also have some major limitations (Chen and Wang, 2014). For the dense systems in where a large number of particles are involved, the calculation of particle–particle interactions is very complex. It is not possible, even with super computers, to simulate a large-scale system due to the extensive computational cost of tracking each particle (Li et al., 2012). Because of this complexity of calculating particle-particle collisions and the high collision frequency for volume fractions above 5%, these calculations have been limited to the order of 2×10^5 particles and are often restricted to two-dimensional solutions without a fluid phase (Snider et al., 2011). To avoid this restriction, some methods have been developed with improvements in calculating particle-particle and particle-wall interactions and also with concept of parcels. The concept of parcel is used to reduce the numbers of particles involved in computations, resulting in a significant acceleration of the speed of simulations (Chen and Wang, 2014). According to (Garg et al., 2012), all publicly available codes except for MFiX-DEM employ a parcel-based approach for the discrete phase. In the parcel approach, a finite number of parcels are tracked rather than using actual individual particles. Each parcel may represent a fractional number of real particles. Typically, several particles with same properties (species, size, density, temperature, etc.) are grouped and put into a parcel. This parcel is also called as a computational/numerical/notional/nominal particle in different literature. However, as ANSYS Fluent mentioned, convergence issues can arise, if fluid volume fraction becomes zero due to either when parcel size is bigger than cell size or too many parcels are squeezed into a cell due to softness of particles. Larger parcel size reduces the number of parcels for a certain mass flow hence lower computational cost. However, the smallest cell should be larger than the largest parcel size (as explained above). Therefore, finding the balance for the optimum mesh is important when using parcel concept. Brief overview of some of Eulerian-Lagrangian models are presented in next sections.

4.1 Lagrangian Discrete Phase Model (DPM)

For low and intermediate solids loading, the inter-particle spacing is high and hence the negligence of particle-particle interactions might be justifiable. The commercial code; ANSYS Fluent has Lagrangian Discrete Phase Model (DPM) with such a treatment for the flows with solids volume fraction less than 10%. In that model, the volume occupied by solids is not taken into account when assembling the continuous phase equations and particle pressure and viscous stresses due to particles are neglected. The fluid carrier influences the particulate phase via drag and turbulence and if the interaction with continuous phase is enabled, additionally the particles in turn influence the carrier fluid via reduction in mean momentum and turbulence.

So, this method has either one-way or two-way coupling between the phases, but not four-way coupling where the particle-particle interactions are considered (Fluent, 2013). The particle-wall collisions are modeled through relatively simple models, often based on a simple reflection coefficients of restitution.

4.2 Dense Discrete Phase Model Incorporated with Kinetic Theory of Granular Flow (DDPM-KTGF)

Dense Discrete Phase Model incorporated with Kinetic Theory of Granular Flow (DDPM-KTGF) for modeling particle-particle and particle-wall interactions are a quite recently developed model. This model is available in commercial code ANSYS Fluent and open source code OpenFOAM®. This is a hybrid model composed with Eulerian-Eulerian and Eulerian-Lagrangian approaches. In low solids volume fractions, the particles are treated in a Lagrangian manner, while in high solids volume fractions, the particles are treated using Eulerian treatment. The solids stress acting on particles resulting from inter-particle interactions is computed from the stress tensor given by the KTGF which is similar to Eulerian-Eulerian approach for granular flows (Euler-granular model). Compared to Lagrangian DPM, this model extends the applicability from dilute to dense phase since this accounts for the effect of volume fraction of solid phase and particle-particle interactions. Still the preciseness of treating particle-particle interactions with KTGF is doubtful. Despite, having benefits of Lagrangian methods and is applicable to large systems, it demands further tests and validations. Some predictions for coal gasification and coal oxy-fuel combustion in circulating fluidized beds (Adamczyk et al., 2014a; Klimanek et al., 2015), circulating fluidized bed boiler (Adamczyk et al., 2014b), impinging particle jet in a channel (Chen and Wang, 2014), solid sorbent carbon capture reactor (Ryan et al., 2013) and ceramic dispersion in liquid pool (Zhang and Nastac, 2014) are made using DDPM-KTGF model. (Ryan et al., 2013) have experienced less stability of DDPM-KTGF solution compared to Euler-granular model and MP-PIC method for a given reactor design and (Chen and Wang, 2014) highlights the requirement of further improvements for DDPM-KTGF model.

4.3 CFD-Discrete Element Method (CFD-DEM)

Soft sphere model based on Cundall and Strack, also called “Discrete Element Method (DEM)” or “Distinct Element Method” can be used to explicitly track the particle-particle and particle-wall interaction terms in typical Eulerian-Lagrangian approach (Crowe et al., 2012). This model approach is often referred to as “CFD-DEM” in most of the literature. In-house developed CFD-DEM codes or DEM codes coupled to available CFD platforms through user defined functions are quite common practices. Standalone DEM simulation codes (codes for pure particulate flows without carrier fluid) include open source codes, such as LAMMPS and YADE, and commercial codes, such as EDEM® and ITASCA. Efforts to couple such standalone DEM codes to existing computational fluid dynamic solvers have recently been undertaken. For example, the EDEM code provides users the ability to couple its DEM modules with other CFD codes such as ANSYS Fluent. Recently, OpenFOAM® has been coupled to YADE and LAMMPS (Garg et al., 2012). In DEM, the whole process of collision or contact is solved by numerical integration of the equations of motion. A collision is treated as a continuous process that occurs over a finite time wherein the contact force is calculated as a continuous function of the distance between colliding particles. These are based on physically realistic interaction laws; as example spring, spring dashpot and Coulomb’s law of friction. Empirical values for the spring stiffness coefficient, damping constant and friction coefficient are required. Compared to Lagrangian DPM, this model gives more accurate predictions for dense and near-packing limit, however at the cost of slower computations. As many other Eulerian-Lagrangian models, CFD-DEM incorporates with parcel concept in some codes, since recently. The parcel concept reduces an inherent limitation of using DEM in large-scale and dense particle systems. Explicitly tracking collisions of all real particles demands very high computational cost compared to tracking parcels which consist of group of real particles. Billions of real particles in large commercial systems can be analyzed using millions of parcels (Snider, 2007). As example, in-built DEM capability including parcel concept is now available in CFD solver, ANSYS Fluent. It is called Dense Discrete Phase Model incorporated with Discrete Element Method (DDPM-DEM) and this is quite a new feature in ANSYS Fluent. Published data for the application of DDPM-DEM are rare and some information can be found for modeling of micron-particle transport, interactions and deposition in triple lung-airways (Feng and Kleinstreuer, 2014) and coal-direct chemical-looping combustion (Zhang et al., 2014). Another CFD-DEM code; MFiX-DEM is limited to small problem sizes due to high computational cost incurred in the particle neighbor search algorithm in where real particles are considered (Garg et al., 2012). The CFD-DEM has been extensively proven to be effective in many gas-solids applications (Chen and Wang, 2014).

4.4 Computational Particle Fluid Dynamics (CPFD) Numerical Scheme Incorporated with the Multiphase-Particle-in-Cell (MP-PIC) Method

The Computational Particle Fluid Dynamics (CPFD) numerical scheme incorporated with the MultiPhase-Particle-In-Cell (MP-PIC) method to describe the solid phase is quite new Eulerian-Lagrangian approach for calculating gas-solids flows. This is a version after several significant improvements of Particle-In-Cell (PIC) method used for single-phase flows since 1960s (Snider, 2001). As Snider, Clark and O'rourke mentioned, the MP-PIC method is, in turn, an extension of the stochastic particle method of the KIVA code (Snider et al., 2011). In the CPFD method, the real particles are grouped into parcels as in many other Eulerian-Lagrangian methods (Zhang et al., 2012). The dynamics of the particle phase is predicted in the MP-PIC method by solving a transport equation which is called Liouville equation for the particle distribution function. The particle distribution function contains particle properties as example, particle spatial location, particle velocity, particle mass, time, etc. (Karimipour and Pugsley, 2012). Unlike DEM models which calculate particle-to-particle force by a spring–damper model and direct particle contact, the CPFD methodology models particles' collision force on each particle as a spatial gradient. A particle normal stress model is developed from this concept to describe the particle collisions (Wang et al., 2014b). In the computation, the stress gradient on the grid is first calculated and then interpolated to discrete particles (Abbasi et al., 2013). The model has been undergone through many improvements such as including Bhatnager, Gross and Krook (BKG) collision model for gas/liquid/solids flows (O'Rourke et al., 2009), including collision damping fluctuations due to inelastic collisions (O'Rourke and Snider, 2010), including return-to-isotropy term in collision source term (O'Rourke and Snider, 2012), including the effects of the contact force variations caused by inhibition of relative motions due to different particle sizes and densities (O'Rourke and Snider, 2014), etc. Arena-Flow®, Barracuda® and OpenFOAM® are some examples for the software/codes which have CPFD implementation. Compared to Lagrangian DPM, this model can accurately model gas-solids flows of dense and close-pack limits. Solution cost is reduced since the collisions are not directly solved as in DEM and also due to implementation of the parcel concept. Furthermore, MP-PIC method does not need to take the particle collisions implicitly, therefore a much larger time step can be adopted (Yin et al., 2014). As mentioned (Lu et al., 2014), this method can be used to model systems with physical particle counts over 1×10^{15} particles. In addition, the CPFD method has shown the ability to model full particle size distribution for any number of solid species and to model particle volume fraction from dilute (<0.1%) upto dense (>60%). Some of the applications of MP-PIC method are bubbling and circulating fluidized beds (Chen et al., 2013; Jiang et al., 2014; Karimipour and Pugsley, 2012; Lan et al., 2013; Liang et al., 2014; Parker et al., 2013; Wang et al., 2014b; Weber et al., 2013; Yin et al., 2014; Zhang et al., 2012), fluidized bed gasifiers (Abbasi et al., 2011; Loha et al., 2014; Singh et al., 2013; Snider et al., 2011; Thapa et al., 2014), fluidized beds for carbon capture (Breault and Huckaby, 2013; Clark et al., 2013; Parker, 2014; Ryan et al., 2013), gas/liquid/solid fluidized beds (O'Rourke et al., 2009; Vivacqua et al., 2013; Zhao et al., 2009), Rayleigh-Taylor mixing layers (Snider, 2001), sedimentation (Andrews and O'Rourke, 1996; Snider, 2001), downer reactors (Abbasi et al., 2012, 2013), dryer (Bigda, 2014) , 3-D particle jet (Snider, 2001), hopper flow (Lu et al., 2014; Snider, 2007), particle flow in U-tube (Snider, 2007).

In addition to these models, Sommerfeld has developed a stochastic collision model to model the inter-particle collisions (Laín and Sommerfeld, 2012). Furthermore, a brief comparison of results obtained using above mentioned models can be found in elsewhere (Chen and Wang, 2014).

5 Conclusions

A general overview of some of the available gas-solids flow modeling approaches is made in the current review paper. Eulerian-Eulerian and Eulerian-Lagrangian are the approaches in use. Further, Lagrangian Discrete Phase Model (DPM), Dense Discrete Phase Model incorporated with Kinetic Theory of Granular Flow (DDPM-KTGF), CFD-Discrete Element Method (CFD-DEM) and Computational Particle Fluid Dynamics (CPFD) numerical scheme incorporated with the MultiPhase-Particle-In-Cell (MP-PIC) method are the models discussed under Eulerian-Lagrangian approach.

The conventional Eulerian-Eulerian model for granular flows and CFD-DEM models have widely been used for many applications and validated quite well. Despite this, both models still have major limitations with respect to accuracy and computational cost, hence applying to large scale systems and to model flows with different particle properties are not very straightforward. Therefore, these models are being under improvements and some new models have been introduced to model gas-solids flows, as example DDPM-KTGF, DDPM-DEM and MP-PIC. In addition to getting advantage of Lagrangian treatment of the particles, these models are said to be efficient compared to the conventional models. This might be due to the use of parcel concept and/or due to use of empirical approaches for modeling particle-particle interactions, alternative algorithms and grid. Few

publications related to use of MP-PIC method are available mainly in fluidized bed applications, however published information for the applications of other models are not very abundant. Therefore, the applicability and validity of these quite recent models for the accurate predictions of gas-solids multiphase flow modeling should be investigated. Moreover, all the models need further improvements in order to apply for wide range of applications and scales.

Acknowledgements

The authors would like to acknowledge the financial support provided by the Research Council of Norway under PETROMAKS II program and Det Norske oljeselskape ASA.

References

A. Abbasi, P. E. Ege, and H. I. de Lasa. CPFD simulation of a fast fluidized bed steam coal gasifier feeding section. *Chemical Engineering Journal, 174*(1): 341-350, 2011. doi:10.1016/j.cej.2011.07.085.

A. Abbasi, M. A. Islam, P. E. Ege, and H. I. de Lasa. Downer reactor flow measurements using CREC-GS-Optiprobes. *Powder Technology, 224*(Supplement C): 1-11, 2012. doi:10.1016/j.powtec.2012.02.005.

A. Abbasi, M. A. Islam, P. E. Ege, and H. I. de Lasa. CPFD flow pattern simulation in downer reactors. *AIChE Journal, 59*(5): 1635-1647, 2013. doi:10.1002/aic.13956.

W. P. Adamczyk, P. Kozołub, G. Węcel, A. Klimanek, R. A. Białecki, and T. Czakiert. Modeling oxy-fuel combustion in a 3D circulating fluidized bed using the hybrid Euler–Lagrange approach. *Applied Thermal Engineering, 71*(1): 266-275, 2014a. doi:10.1016/j.applthermaleng.2014.06.063.

W. P. Adamczyk, G. Węcel, M. Klajny, P. Kozołub, A. Klimanek, and R. A. Białecki. Modeling of particle transport and combustion phenomena in a large-scale circulating fluidized bed boiler using a hybrid Euler–Lagrange approach. *Particuology, 16*(Supplement C): 29-40, 2014b. doi:10.1016/j.partic.2013.10.007.

M. J. Andrews and P. J. O'Rourke. The multiphase particle-in-cell (MP-PIC) method for dense particulate flows. *International Journal of Multiphase Flow, 22*(2): 379-402, 1996. doi:10.1016/0301-9322(95)00072-0.

H. Arastoopour. Numerical simulation and experimental analysis of gas/solid flow systems: 1999 Fluor-Daniel Plenary lecture. *Powder Technology, 119*(2): 59-67, 2001. doi:10.1016/S0032-5910(00)00417-4.

J. Bigda. CPFD Numerical Study of Impact Dryer Performance. *Drying Technology, 32*(11): 1277-1288, 2014. doi:10.1080/07373937.2014.929586.

R. W. Breault and E. D. Huckaby. Parametric behavior of a CO_2 capture process: CFD simulation of solid-sorbent CO_2 absorption in a riser reactor. *Applied Energy, 112*(Supplement C): 224-234, 2013. doi:10.1016/j.apenergy.2013.06.008.

C. Chen, J. Werther, S. Heinrich, H.-Y. Qi, and E.-U. Hartge. CPFD simulation of circulating fluidized bed risers. *Powder Technology, 235*(Supplement C): 238-247, 2013. doi:10.1016/j.powtec.2012.10.014.

X. Chen and J. Wang. A comparison of two-fluid model, dense discrete particle model and CFD-DEM method for modeling impinging gas–solid flows. *Powder Technology, 254*(Supplement C): 94-102, 2014. doi:10.1016/j.powtec.2013.12.056.

S. Clark, D. M. Snider, and J. Spenik. CO_2 Adsorption loop experiment with Eulerian–Lagrangian simulation. *Powder Technology, 242*(Supplement C): 100-107, 2013. doi:10.1016/j.powtec.2013.01.011.

C. T. Crowe, J. D. Schwarzkopf, M. Sommerfeld, and Y. Tsuji. Multiphase flows with droplets and particles, Taylor & Francis Group, LLC. 2012.

Y. Feng and C. Kleinstreuer. Micron-particle transport, interactions and deposition in triple lung-airway bifurcations using a novel modeling approach. *Journal of Aerosol Science, 71*(Supplement C): 1-15, 2014. doi:10.1016/j.jaerosci.2014.01.003.

ANSYS fluent theory guide 15.0. Canonsburg, PA, ANSYS Inc. 2013.

R. Garg, J. Galvin, T. Li, and S. Pannala. Open-source MFIX-DEM software for gas–solids flows: Part I—Verification studies. *Powder Technology, 220*(Supplement C): 122-137, 2012. doi:10.1016/j.powtec.2011.09.019.

Y. Jiang, G. Qiu, and H. Wang. Modelling and experimental investigation of the full-loop gas–solid flow in a circulating fluidized bed with six cyclone separators. *Chemical Engineering Science, 109*(Supplement C): 85-97, 2014. doi:10.1016/j.ces.2014.01.029.

S. Karimipour and T. Pugsley. Application of the particle in cell approach for the simulation of bubbling fluidized beds of Geldart A particles. *Powder Technology, 220*(Supplement C): 63-69, 2012. doi:10.1016/j.powtec.2011.09.026.

A. Klimanek, W. Adamczyk, A. Katelbach-Woźniak, G. Węcel, and A. Szlęk. Towards a hybrid Eulerian–Lagrangian CFD modeling of coal gasification in a circulating fluidized bed reactor. *Fuel, 152*(Supplement C): 131-137, 2015. doi:10.1016/j.fuel.2014.10.058.

S. Laín and M. Sommerfeld. Numerical calculation of pneumatic conveying in horizontal channels and pipes: Detailed analysis of conveying behaviour. *International Journal of Multiphase Flow, 39*(Supplement C): 105-120, 2012. doi:10.1016/j.ijmultiphaseflow.2011.09.006.

X. Lan, X. Shi, Y. Zhang, Y. Wang, C. Xu, and J. Gao. Solids Back-mixing Behavior and Effect of the Mesoscale Structure in CFB Risers. *Industrial & Engineering Chemistry Research, 52*(34): 11888-11896, 2013. doi:10.1021/ie3034448.

T. Li, R. Garg, J. Galvin, and S. Pannala. Open-source MFIX-DEM software for gas-solids flows: Part II — Validation studies. *Powder Technology, 220*(Supplement C): 138-150, 2012. doi:10.1016/j.powtec.2011.09.020.

Y. Liang, Y. Zhang, T. Li, and C. Lu. A critical validation study on CPFD model in simulating gas–solid bubbling fluidized beds. *Powder Technology, 263*(Supplement C): 121-134, 2014. doi:10.1016/j.powtec.2014.05.003.

C. Loha, H. Chattopadhyay, and P. K. Chatterjee. Three dimensional kinetic modeling of fluidized bed biomass

gasification. *Chemical Engineering Science, 109*(Supplement C): 53-64, 2014. doi:10.1016/j.ces.2014.01.017.

H. Lu, X. Guo, W. Zhao, X. Gong, and J. Lu. Experimental and CPFD Numerical Study on Hopper Discharge. *Industrial & Engineering Chemistry Research, 53*(30): 12160-12169, 2014. doi:10.1021/ie403862f.

P. J. O'Rourke and D. M. Snider. Inclusion of collisional return-to-isotropy in the MP-PIC method. *Chemical Engineering Science, 80*(Supplement C): 39-54, 2012. doi:10.1016/j.ces.2012.05.047.

P. J. O'Rourke and D. M. Snider. A new blended acceleration model for the particle contact forces induced by an interstitial fluid in dense particle/fluid flows. *Powder Technology, 256*(Supplement C): 39-51, 2014. doi:10.1016/j.powtec.2014.01.084.

P. J. O'Rourke and D. M. Snider. An improved collision damping time for MP-PIC calculations of dense particle flows with applications to polydisperse sedimenting beds and colliding particle jets. *Chemical Engineering Science, 65*(22): 6014-6028, 2010. doi:10.1016/j.ces.2010.08.032.

P. J. O'Rourke, P. Zhao, and D. Snider. A model for collisional exchange in gas/liquid/solid fluidized beds. *Chemical Engineering Science, 64*(8): 1784-1797, 2009. doi:10.1016/j.ces.2008.12.014.

J. Parker, K. LaMarche, W. Chen, K. Williams, H. Stamato, and S. Thibault. CFD simulations for prediction of scaling effects in pharmaceutical fluidized bed processors at three scales. *Powder Technology, 235*(Supplement C): 115-120, 2013. doi:10.1016/j.powtec.2012.09.021.

J. M. Parker. CFD model for the simulation of chemical looping combustion. *Powder Technology, 265*(Supplement C): 47-53, 2014. doi:10.1016/j.powtec.2014.01.027.

E. M. Ryan, D. DeCroix, R. Breault, W. Xu, E. D. Huckaby, K. Saha, S. Dartevelle, and X. Sun. Multi-phase CFD modeling of solid sorbent carbon capture system. *Powder Technology, 242*(Supplement C): 117-134, 2013. doi:10.1016/j.powtec.2013.01.009.

R. I. Singh, A. Brink, and M. Hupa. CFD modeling to study fluidized bed combustion and gasification. *Applied Thermal Engineering, 52*(2): 585-614, 2013. doi:10.1016/j.applthermaleng.2012.12.017.

D. M. Snider. An Incompressible Three-Dimensional Multiphase Particle-in-Cell Model for Dense Particle Flows. *Journal of Computational Physics, 170*(2): 523-549, 2001. doi:10.1006/jcph.2001.6747.

D. M. Snider. Three fundamental granular flow experiments and CPFD predictions. *Powder Technology, 176*(1): 36-46, 2007. doi:10.1016/j.powtec.2007.01.032.

D. M. Snider, S. M. Clark, and P. J. O'Rourke. Eulerian–Lagrangian method for three-dimensional thermal reacting flow with application to coal gasifiers. *Chemical Engineering Science, 66*(6): 1285-1295, 2011. doi:10.1016/j.ces.2010.12.042.

R. Thapa, C. Pfeifer, and B. Halvorsen. Modeling of reaction kinetics in bubbling fluidized bed biomass gasification reactor. *Internal Journal of Energy and Environment, 5*(1): 35-44, 2014.

V. Vivacqua, S. Vashisth, A. Prams, G. Hébrard, N. Epstein, and J. R. Grace. Experimental and CPFD study of axial and radial liquid mixing in water-fluidized beds of two solids exhibiting layer inversion. *Chemical Engineering Science, 95*(Supplement C): 119-127, 2013. doi:10.1016/j.ces.2013.03.011.

Q. Wang, H. Yang, P. Wang, J. Lu, Q. Liu, H. Zhang, L. Wei, and M. Zhang. Application of CPFD method in the simulation of a circulating fluidized bed with a loop seal Part II—Investigation of solids circulation. *Powder Technology, 253*(Supplement C): 822-828, 2014a. doi:10.1016/j.powtec.2013.11.040

Q. Wang, H. Yang, P. Wang, J. Lu, Q. Liu, H. Zhang, L. Wei, and M. Zhang. Application of CPFD method in the simulation of a circulating fluidized bed with a loop seal, part I—Determination of modeling parameters. *Powder Technology, 253*(Supplement C): 814-821, 2014b. doi:10.1016/j.powtec.2013.11.041.

J. M. Weber, K. J. Layfield, D. T. Van Essendelft, and J. S. Mei. Fluid bed characterization using Electrical Capacitance Volume Tomography (ECVT), compared to CPFD Software's Barracuda. *Powder Technology, 250*(Supplement C): 138-146, 2013. doi:10.1016/j.powtec.2013.10.005.

S. Yin, W. Zhong, B. Jin, and J. Fan. Modeling on the hydrodynamics of pressurized high-flux circulating fluidized beds (PHFCFBs) by Eulerian–Lagrangian approach. *Powder Technology, 259*(Supplement C): 52-64, 2014. doi:10.1016/j.powtec.2014.03.059.

D. Zhang and L. Nastac. Numerical modeling of the dispersion of ceramic nanoparticles during ultrasonic processing of aluminum-based nanocomposites. *Journal of Materials Research and Technology, 3*(4): 296-302, 2014. doi:10.1016/j.jmrt.2014.09.001.

Y. Zhang, X. Lan, and J. Gao. Modeling of gas-solid flow in a CFB riser based on computational particle fluid dynamics. *Petroleum Science, 9*(4): 535-543, 2012. doi:10.1007/s12182-012-0240-7.

Z. Zhang, L. Zhou, and R. Agarwal. Transient Simulations of Spouted Fluidized Bed for Coal-Direct Chemical Looping Combustion. *Energy & Fuels, 28*(2): 1548-1560, 2014. doi:10.1021/ef402521x.

P. Zhao, P. J. O'Rourke, and D. Snider. Three-dimensional simulation of liquid injection, film formation and transport, in fluidized beds. *Particuology, 7*(5): 337-346, 2009. doi:10.1016/j.partic.2009.07.002.

A Simulation Model Validation and Calibration Platform

Shenglin Lin, Wei Li, Xiaochao Qian, Ping Ma, Ming Yang*

Control and Simulation Center, Harbin Institute of Technology, China
lin_44627079@yeah.net, frank@hit.edu.cn, everqxc@hotmail.com, pingma@hit.edu.cn, myang@hit.edu.cn

Abstract

The simulation model validation and calibration (SMVC) is a complicated work, including uncertainty description, many simulation experiments execution and complex data analysis etc. Moreover, there are many uncertainty factors such as model form assumptions and solution approximations, random variability of model inputs, etc. need to be considered when requiring a precise model. For assisting the SMVC effectively, this paper develops a software platform to validate and calibrate the simulation models when some quantities may be affected by uncertainty. First, an unprecedented process model, which includes uncertainty description, simulation experiment design, model validation and model calibration, is presented to explicate the procedures of SMVC under uncertainty. In the process model, many new model validation and calibration algorithms under uncertainty are applied based on our previous work. Second, the design of platform is divided into two parts, which consist of structure design and function design, and the software technique based on strategy pattern is introduced to integrate and maintain the SMVC algorithms. Then this platform is implemented according to its expected uses and key design requirements. Finally, application example of model validation and calibration of a flight vehicle kinematic control system is illustrated how to use the platform.

Keywords: model validation, model calibration, uncertainty description, software platform

1 Introduction

Simulation models are increasingly used to solve practical problems in various engineering disciplines. The basic premise of the simulation-based solutions is the credible simulation models, so model validation is introduced naturally, which is used to validate the simulation models by measuring the extent of agreement between the model output and experimental observations (Sankararaman and Mahadevan, 2011). However, the credibility of simulation models are also affected by various sources of uncertainty such as model form assumptions and solution approximations, natural variability in model inputs and parameters, and data uncertainty due to sparse and imprecise information. For improving the precision of simulation models further, the uncertainty parameters must be described and determined. So model calibration is proposed to improve the quality of simulation models through adjusting model parameters according to the results of model validation. Besides, in order to research the SMVC under uncertainty better, the model uncertainties are classified according to their fundamental essence as either (a) aleatory - the inherent variation in a quantity that, given sufficient samples of the stochastic process, can be characterized via a probability density distribution, or (b) epistemic - uncertainty due to lack of knowledge by the modelers, analysts conducting the analysis, or experimentalists involved in modeling and simulation (Roy and Oberkampf, 2011).

The simulation model validation and calibration (SMVC) process under uncertainty is a complicated and tedious work, which involves quantification of various uncertainties, design of many simulation experiments, complex data management and analysis, application of many kinds of SMVC methods, etc. Thus, an auxiliary platform is needed to develop for assisting the SMVC under uncertainty. Currently, the development of SMVC under uncertainty platform is still in the initial stage, such as some modeling and validation platforms are listed below. A comprehensive platform of modeling and simulation credibility evaluation is developed (Balci *et al*, 2002). An integrated simulation measurement platform is developed by Institute for Simulation and Training (IST). Besides, some tools about uncertainty quantification and analysis are exploited. For example, an uncertainty quantification and analysis tool is implemented for the automated operation of uncertainty quantification and analysis (Ferson, 2002). An automated tool about the probability boundary calculation interval-based is developed by Berleant (1993). Simlab is an integrated uncertainty and sensitivity analysis tool, which is used to solve the sample generation of uncertainty variables, uncertainty propagation and analysis (Saltelli *et al*, 2005).

The existing auxiliary platforms above cannot accomplish the SMVC under certainty completely and the integrated auxiliary platforms have not been researched. In order to assist to validate and calibrate simulation models, an integrated software platform, HIT-MVCP (Harbin Institute of Technology Model Validation and Calibration Platform), is designed and implemented, which can assist users to accomplish SMVC under uncertainty effectively. The remainder of

this paper is organized as follows. A process model for SMVC under uncertainty is presented in Section 2. The structure and function of HIT-MVCP are designed in Section 3. Section 4 describes the software implementation of HIT-MVCP and an example of model validation and calibration of a flight vehicle kinematic control system is illustrated how to use the platform. Finally, the conclusion and the future work are summarized.

2 The Process Model of SMVC under Uncertainty

Due to the complex and tedious model validation and calibration procedures, a reasonable and explicit operation scheme is necessary before implementing SMVC under uncertainty. So a process model of SMVC under uncertainty is presented, which includes uncertainty description, experiment schemes design, simulation execution, model validation, consistency metamodeling, uncertainty parameters optimization (Figure 1). First, various uncertainty variables which impact the simulation execution should be determined and described primarily in special forms such as stochastic and interval variable (Helton, 2011). After that, for the subsequent model calibration according to the validation results, the propagation of the uncertainty effect from input information to simulation results needs to be researched (Helton *et al*, 2006) and the simulation experiment design is naturally used to solve this problem based on the design purposes and uncertainty expression. In the next step, the credibility of simulation model could be achieved through measuring the extent of agreement between the model output and experimental observations (Meng *et al*, 2015). Finally, for accomplishing model calibration, we merely need to adjust repeatedly the uncertainty parameters to maximize the data consistency. But, due to the effect of aleatory and epistemic uncertainty, the calibration process of uncertainty parameters needs plenty of experiment schemes design and simulation execution times and this will produce much computation cost. So a metamodel-based algorithm is applied to improve the efficiency of model calibration and reduce the computation expense. The specific description of the algorithm is researched in (Qian *et al*, 2016).

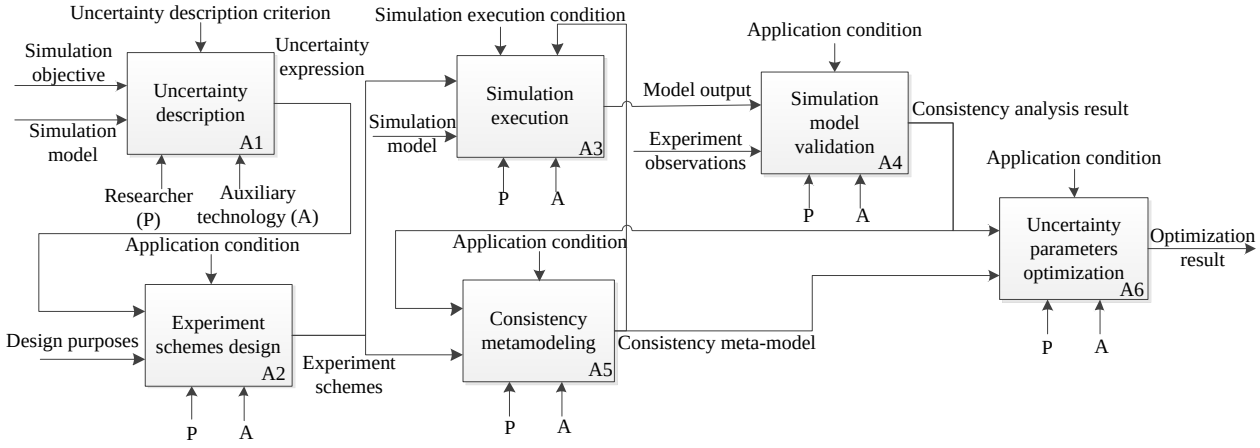


Figure 1. The SMVC process model.

The procedures of SMVC under uncertainty include following parts:

A. Uncertainty description (A1): According to the simulation objective, the uncertainty information involved in simulation execution would be described. The aleatory and epistemic uncertainties are expressed respectively in form of stochastic variable and interval variable.

B. Experiment schemes design and simulation execution (A2, A3): The uncertainty information is propagated from input parameters to simulation results based on the experiment design. The probability sampling methods are used such as, random sampling, LHS sampling, uniform sampling etc. and the experiment schemes could be stored and reused.

C. Simulation model validation (A4): We need to validate the extent of agreement between the model output and the experiment observations. These data with the simulation uncertainty often belong to different styles, such as dynamic data, static data etc. So, a validation method based evidence distance for multi-outputs under uncertainty is introduced (Qian *et al*, 2016) and the brief procedures are show as follows.

a. Extracting the data feature, such as shape, position, frequency spectrum etc. And constructing the feature matrix.

b. Evidence fusion. Transferring the feature matrix into form of evidence body.

c. The validation result could be achieved by means of computing the evidence distance among the evidence bodies.

D. Consistency metamodeling (A5): For reducing the

computation expense, the consistency metamodel of multiple outputs is constructed. The simulation model calibration would be completed via model calibration and uncertainty parameters optimization. The epistemic uncertainty parameters are sampled by means of the experiment design and the simulation experimental outputs under each epistemic uncertainty sample are obtained. Each simulation result which only involves the aleatory uncertainty is used to make a consistency analysis with the reference data, and then the agreement metamodel with the aleatory uncertainty variable could be obtained.

E. Uncertainty parameters optimization (A6): The consistency metamodel is regarded as the objective function in the optimization process of epistemic uncertainty parameters and many optimization algorithms are used to achieve the appropriate parameter value. The procedures of simulation model calibration based on consistency metamodeling and parameters optimization are given as follows:

a. Designing and executing simulation experiment schemes to obtain the epistemic uncertainty samples and simulation results.

b. Evaluating the agreement degree between the model output and experiment observations.

c. According to the multiple epistemic uncertainty samples and the results of model validation, the consistency metamodel could be constructed.

d. To maximize the data consistency, the optimal epistemic uncertainty parameter could be determined.

3 The Structure and Function Design of HIT-MVCP

The proposed process model in Section 2 is regarded as the guideline of SMVC under uncertainty. According to the function description of each procedure in the process model, an auxiliary platform, HIT-MVCP, is designed which involves structure design and function design. However, many developing SMVC algorithms under uncertainty such as methods of the experiment scheme design, model validation algorithms, metamodeling methods etc. need to be integrated and maintained in the design process. If we maintain the platform continuously with the fast development of SMVC algorithms and this will generate high expense, and reduce the practicality of platform. So a software technique based on strategy pattern is introduced to design and implement the algorithms involved in SMVC process, and improve the maintainability and extensibility of HIT-MVCP.

3.1 The Structural Design of HIT-MVCP

Based on the functions and usage description of each procedure in the SMVC process model and modularized principle, the structure of HIT-MVCP is divided into five subsystems to respectively design, which including uncertainty description, experiment design and execution, simulation model validation, simulation model calibration, data management.

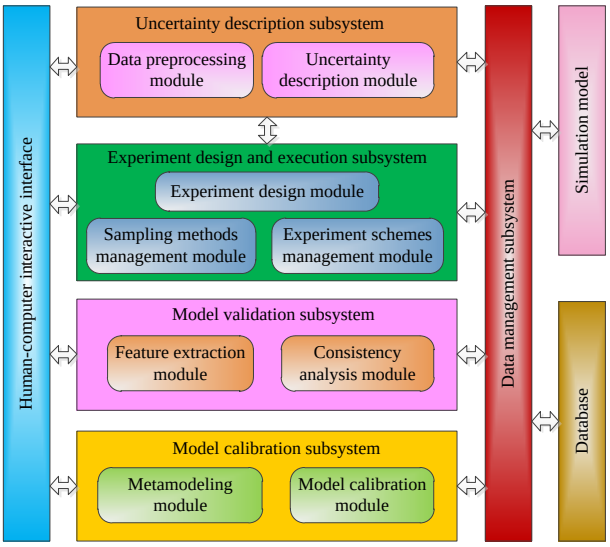


Figure 2. The structure design of HIT-MVCP.

The detailed design described in Figure 2 includes following subsystems:

A. Uncertainty description subsystem. In the subsystem of uncertainty description, we need to determine uncertainty factors of the simulation models and describe them in forms of probability based on the experiment observations and model output. Due to the uncertainty effect, there are some defects such as outliers etc. in simulation results. So the data preprocessing procedure need to be introduced and then the uncertainty variables can be expressed based on the data post-processing.

B. Simulation experiment design and execution subsystem. This subsystem consists of three modules, which include experiment design, sampling methods management and experiment schemes management. The experiment design module is used to produce experiment schemes based on the sampling method from the module of sampling methods management. The experiment schemes are managed in experiment schemes management module and transmitted to data management subsystem. Then, the simulation model will receive and execute these experiment schemes.

C. Model validation subsystem. This subsystem is used to measure the degree of agreement between the model output and experimental observations. Based on the introduced model validation algorithm, the model validation subsystem is divided into two modules to design respectively, which include feature extraction and consistency analysis. And then the validation results will be stored in database through the data management

subsystem.

D. Model calibration subsystem. Based on the presented calibration algorithm in Section 2, the model calibration subsystem is divided into metamodeling and model calibration. The consistency metamodel under aleatory uncertainty is constructed based on the experiment schemes and results of model validation. The consistency metamodel is regard as the objective function and the model calibration module is used to adjust epistemic uncertainty parameter based on many optimization algorithms.

E. Data management subsystem. This subsystem is responsible for providing the interaction interface between simulation model and databases. The information of SMVC process such as the storage and execution of experiment schemes, transmission of model output etc. are all managed via this subsystem.

With the development of SMVC techniques, many kinds of algorithms about SMVC under uncertainty are researched and applied such as random sampling, Kriging, Hypothesis testing etc. Assuming we integrate the new algorithms into HIT-MVCP constantly and the high cost and the bad practicability will be arisen. So the software technique of strategy pattern is used to design the related algorithms of SMVC under uncertainty and then the maintainability and extensibility of HIT-MVCP is improved effectively. Strategy pattern makes the alteration of algorithms independent from users through defining and encapsulating a family of algorithms which are interchangeable. The detail algorithm design based on strategy pattern is described in Section B.

According to the process model of SMVC under uncertainty and structure design, the related algorithms of HIT-MVCP are divided into eight types: data preprocessing, uncertainties described, sampling methods, feature extraction, validation principle, calibration principle, optimization methods and metamodeing (Figure 3). These algorithms are used in different phases of SMVC under uncertainty and the extended interface is design based on strategy pattern.

3.2 The Structural Design of HIT-MVCP

Depending on the process model of SMVC under uncertainty, HIT-MVCP is developed to realize two functions: model validation and model calibration. Model validation is used to measure the extent of agreement between model output and experiment observations. Based on the algorithm description of model validation and strategy pattern, model validation function is divided into two parts to design respectively, which including feature extraction and consistency analysis and the class diagram of software design is given in Figure 4. As the core of model validation subsystem, the model validation class relies on the data management subsystem, and helps coordinate the feature extraction and consistency analysis. The interaction procedure of each class is given in Figure 5.

Model calibration is used to determine the epistemic uncertainty of model for maximizing the degree of agreement between model output and experimental observations. According to the algorithm description of model calibration and strategy pattern in Section 2, this subsystem is divided into two modules to design respectively, which includes metamodeling and model calibration. The metamodeling module includes sequential design criterion and metamodeling type class, and completes the metamodeling based on sequential design. The model calibration subsystem involves optimization methods class and calibration criterion class, and is responsible for adjusting the epistemic uncertainty parameter. The class diagram of model calibration subsystem is given in Figure 6. As the core of the subsystem, the model calibration class is used for the data exchange and control, and depends on the data management subsystem to call the experiment schemes. The interaction procedure of each class is given in Figure 7.

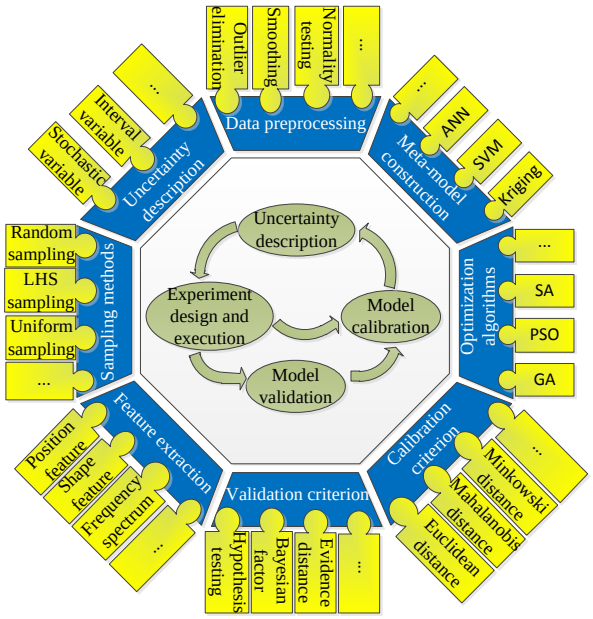


Figure 3. The algorithm structure design.

4 The Implementation and Application of HIT-MVCP

HIT-MVCP is implemented based on C++ language depending on the structure and function design above. The human-computer interaction interface of platform is given in Figure 8 and the all subsystems could be called from this interface. In addition, an example about the model validation and calibration of a flight vehicle kinematic control system is given to validate the efficiency of the described functions and illustrate how to use HIT-MVCP. This kinematic control model is crucial for movement simulation of a flight vehicle.

Multiple disturbances and uncertainties will be encountered in simulation phase, and impact the control accuracy of flight vehicle such as initial mass and velocity of flight vehicle, atmospheric density, damping factor, etc. For making the terminal guidance simulation model more precise, the disturbances and uncertainties affecting terminal guidance model need to be described accurately, and researched how to influence the model prediction. Besides, for evaluating whether the model could reach the application standard of simulation and application, the model credibility under multiple uncertainties must be validated.

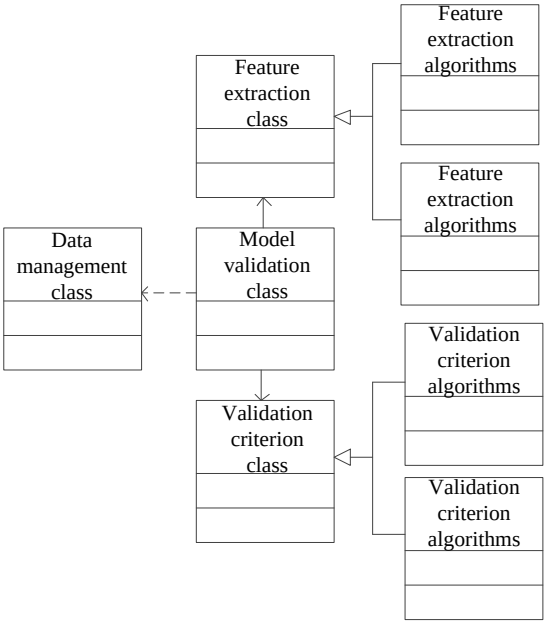


Figure 4. Class diagram of the model validation subsystem.

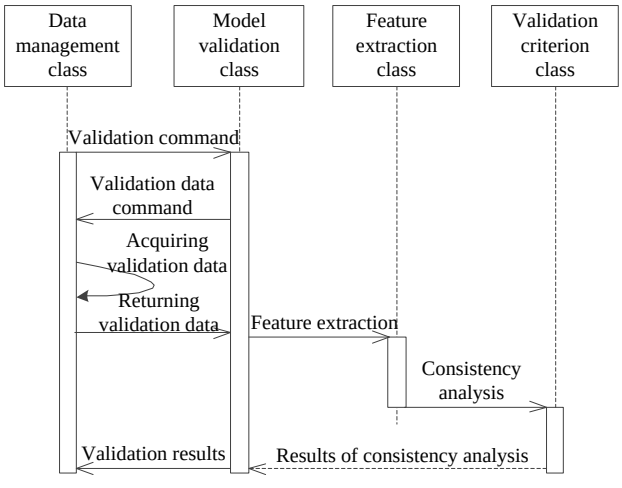


Figure 5. Interactive procedure of model validation.

According to the process model of SMVC under uncertainty, the uncertainty parameters involved in the simulation model firstly need to be described. Supposing the aleatory uncertainties consist of atmospheric density, initial velocity and angle of the flight vehicle, lift coefficient, damping factor and the epistemic uncertainties involve initial mass and reference area of the fight vehicle. The expression manner of the uncertainty parameters could be determined by means of parameter estimation of multiple sample data and user input directly (Figure 9).

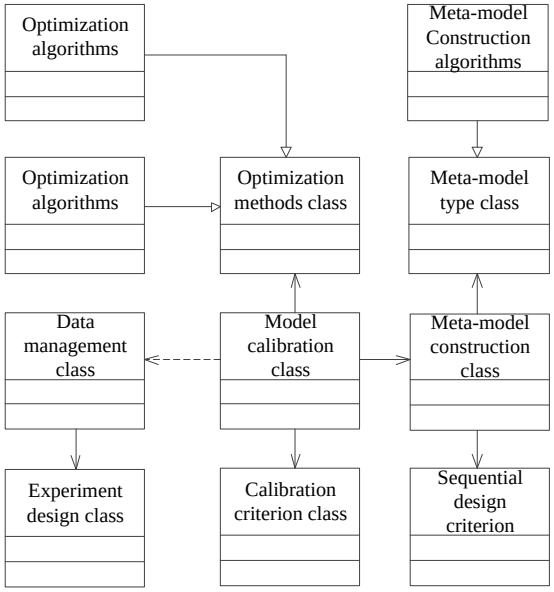


Figure 6. Class diagram of the model calibration subsystem.

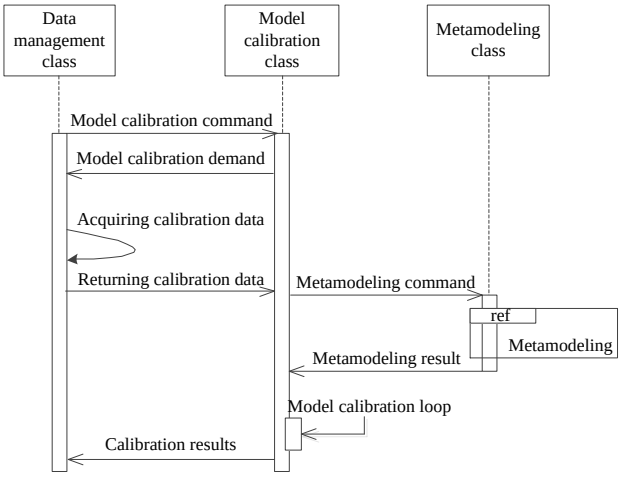


Figure 7. Interactive procedure of model calibration.

After that, the experiment schemes are designed and executed to propagate the uncertainty influence from input parameters to simulation results. LHS method is used to sample the uncertainty parameters. The simulation results could be achieved and used to validate the model credibility with the experimental observations. The result of model validation is given in Figure 10. The blue envelope lines represent the evidence body of model output and the red envelope lines are the evidence body of experiment observations. The evidence distance of two evidence bodies is

0.6581, which represents the model validation result before calibration.

Finally, the epistemic uncertainty parameters of model need to be calibrated. The consistency metamodel which is obtained through multiple model output and experiment observations is established based on sequential design method in Figure 11. The model calibration information are configured such as optimization methods, terminal conditions etc. After calibrating the epistemic uncertainty parameter, the validation result after calibration is given in Figure 12 and the later evidence distance is 0.9398. Through comparing the validation result of before and after calibration, the kinematic model becomes more precise and HIT-MVCP can implement validation and calibration of simulation models effectively.

Figure 8. The human-computer interactive interface of HIT-MVCP.

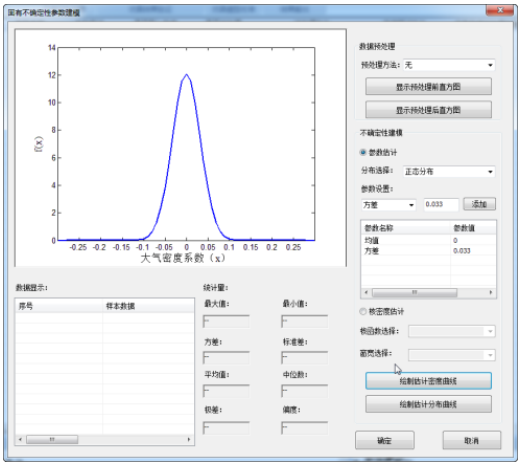


Figure 9. Configuration interface of the uncertainty parameters.

5 Conclusions

With the development of modeling and simulation techniques, more and more complex simulation models need to be validated and calibrated. Due to SMVC under uncertainty is a complicated and tedious work, which includes quantification of various uncertainties, design of many simulation experiments, complex data management and analysis etc. So this paper develops an auxiliary SMVC platform, HIT-MVCP. A process model is presented firstly to explicate the operation procedures of SMVC under uncertainty. Then, the structure and function of HIT-MVCP are designed and the software technique of strategy pattern is introduced to design maintainability and extendibility of the SMVC algorithms. Finally, the software implementation of HIT-MVCP is given and an example of model validation and calibration of a flight vehicle kinematic control system is illustrated how to use the platform. This example shows that HIT-MVCP could assist professional to accomplish the SMVC under uncertainty effectively.

Currently, the epistemic uncertainty is only described based on interval theory and this will cause the incomplete description of uncertainty information sometimes. In future work, many other description methods such as evidence theory and imprecise probability theory etc. need to be introduced to describe the uncertainty parameter. Besides, the SMVC algorithms involved in the platform such as methods of

experiment schemes design, model validation algorithms and model calibration methods etc. are insufficient to adapt to the SMVC of different complex simulation models, so more related algorithms need to be introduced into HIT-MVCP in the future.

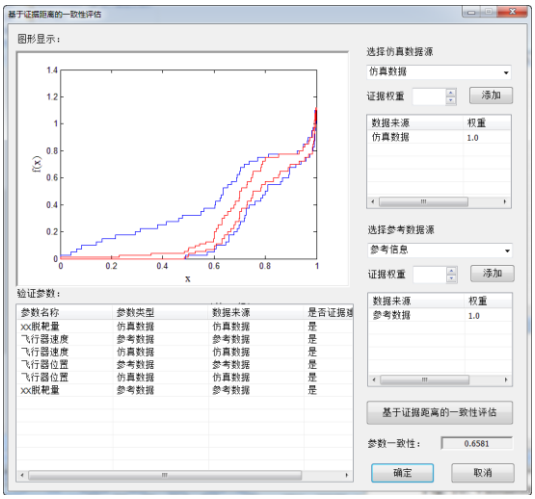


Figure 10. Validation results before calibration.

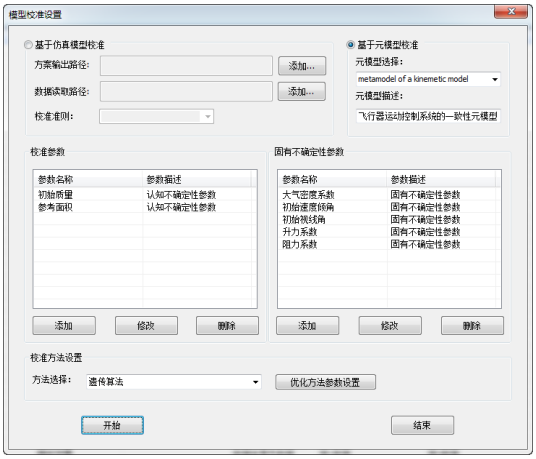


Figure 11. Configuration interface of the model calibration.

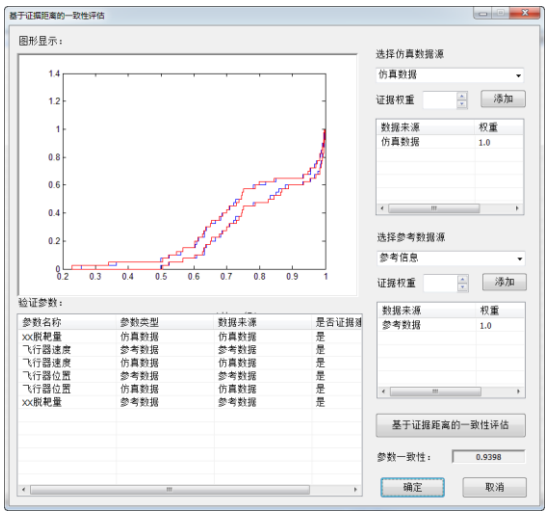


Figure 12. The validation results after calibration.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (Grant No. 61403097).

References

O. Balci, R. J. Adams, and D. S. Myers. A Collaborative Evaluation Environment for Credibility Assessment of Modeling and Simulation Applications. *Simulation Conference, Proceedings of the Winter*. U.S.: San Diego, CA. 214-220, 2002. doi:10.1109/wsc.2002.1172887.

D. Berleant. Automatically Verified Reasoning with both Intervals and Probability Density Functions. *Interval Computations*, 2:48-70, 1993. doi:10.1007/978-1-4613-3440-8_10.

S. Ferson. *RAMAS Risk Calc 4.0 Software: Risk Assessment with Uncertain Numbers*. CRC Press, 2002.

Jon C. Helton. Quantification of Margins and Uncertainties: Conceptual and Computational Basis. *Reliability Engeering & System Safety*, 96:976-1013, 2011. doi:10.1016/j.ress.2011.03.017.

Jon C. Helton, J. D. Johnson, and C. J. Sallaberry. Survey of Sampling-based Methods for Uncertainty and Sensitivity Analysis. *Reliability Engineering & System Safety*, 91:1175-1209. 2006. doi:10.1016/j.ress.2005.11.017.

L. Meng, X. Qian, and H. Wang. Result Validation of A Rudder Simulation Model. *27th European Modeling and Simulation Symposium. Italy: Bergeggi*, 63-69, 2015.

X. Qian, W. Li, and M. Yang. Two-stage Nested Optimization-based Uncertainty Propagation Method for Model Calibration. *International Journal of Modeling, Simulation, and Scientific Computing*, 7:1-17, 2016. doi:10.1007/978-3-642-45037-2_23.

Christopher J. Roy and William L. Oberkampf. A Comprehensive Framework for Verification, Validation, and Uncertainty Quantification in Scientific Computing. *Comput. Methods Appl. Mech. Engrg.* 200:2131-2144, 2011. doi:10.1016/j.cma.2011.03.016.

A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. *Journal of the Royal Statistical Society. Series A, Statistics in Society*, 168:4466-4473, 2005. doi:10.1111/j.1467-985X.2005.358_16.x.

S. Sankararaman and S. Mahadevan. Model Validation under Epistemic Uncertainty. *Reliability Engineering and System Safety*, 96:1232-1241, 2011. doi:10.1016/j.ress.2010.07.014.

The Application of Inflow Control Device for an Improved Oil Recovery using ECLIPSE

Ambrose A. Ugwu Britt M.E Moldestad

Department of Process, Energy and Environmental Technology, University College of Southeast Norway,
britt.moldestad@usn.no

Abstract

The rate of inflow to a horizontal well could vary along the completion length due frictional pressure losses or heterogeneity in the reservoir. These variations reduce oil sweep efficiency and the ultimate recovery. Owing to this, it is necessary to manage fluid flow through the reservoir in order to maximize oil recovery along horizontal wells. One increasingly popular approach is to use inflow control devices (ICD) that delay water and gas breakthrough into the well. Inflow control devices balance the inflow coming from the reservoir towards the wellbore by introducing an extra pressure drop. This paper presents the mathematical models used for the implementation of ICD in ECLIPSE. A case using heterogeneous reservoir similar to Troll offshore Norway was illustrated. The simulation result shows that ICD could delay water breakthrough for 262days and water cut after 3000days reduced by 11%. Gas breakgthrough was also reduced by approximately 51% with ICD.

Keywords: ECLIPSE, IOR, ICD, inflow

1 Introduction

The challenges introduced by reservoir heterogeneity with horizontal wells tend to increase with increasing well length (Birchenko et al, 2011). Completions with long intervals often have significantly uneven specific inflow distribution along their length. These inflow variations cause premature water or gas breakthrough and should be minimized (Hallundbæk and Hazel, 2016). Advanced well completions have been demonstrated as solution to these challenges. Inflow Control Devices (ICDs) is an established type of advanced completions that provide passive inflow control (Henriksen et al, 2006). ICDs are widely used and can be considered to be a mature well completion technology. One of the challenges is the variation in rock properties. Figure 1 illustrates a typical orifice ICD.

Fluid specific inflow rate tends to increase with increasing well length (Krinis et al, 2009). The performance of ICDs can be analyzed in detail with the help of various reservoir simulation tools such as ECLIPSE (Birchenko et al, 2011). ECLIPSE includes basic functionality for ICD modeling (Birchenko et al, 2011) and also offers a practical means to capture the effect of annular flow. ICDs are static and usually installed at the beginning of the production life. An alternative technology is the use of autonomous inflow control device with the ability of closing off the flow interval in an event of water or gas breakthrough (Birchenko et al, 2011).

This paper presents ECLIPSE model for the application of ICD in heterogeneous reservoirs. From the mathematical models, the parameters that substantially reduce the inflow variation can be determined. A case study was simulated to illustrate the impact of a specific ICD completion on Inflow performance at Troll offshore Norway.

2 ECLIPSE Computational Model

In ECLIPSE, ICD is used to control the inflow profile along a horizontal well or branch by imposing an additional pressure drop between the sand face and the tubing. The device is placed around a section of the tubing and diverts the fluid inflowing from the adjacent part of the formation through a sand screen and then into a spiral before it enters the tubing (Mathiesen et al, 2011).

2.1 Pressure drop

The pressure drop across the device is calculated from calibration data, adjusted to allow for the varying density and viscosity of the reservoir fluid flowing

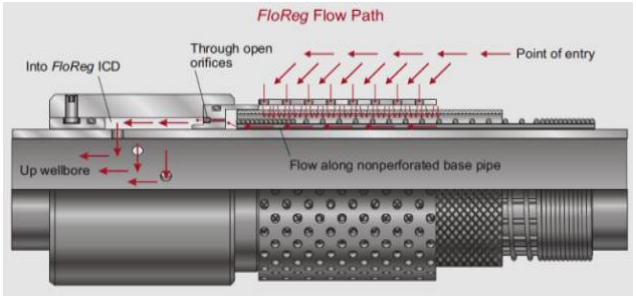


Figure 1. Oriface ICD (Birchenko et al, 2011).

through the device. The pressure drop equation is shown in (1) below (Schlumberger, 2013).

$$\partial P = \left(\frac{\rho_{cal}}{\rho_{mix}} \cdot \frac{\mu_{mix}}{\mu_{cal}}\right)^{1/4} \cdot \frac{\rho_{mix}}{\rho_{cal}} \cdot K \cdot q^2) \quad (1)$$

Here ρmix is the density of the fluid mixture in the segment at local conditions and ρcal is the density of the fluid used to calibrate the ICD. μmix is the viscosity of the fluid mixture in the segment at local conditions and μmix is the viscosity of the fluid used to calibrate the ICD. K is the base strength of the ICD defined in (2).

$$K = \frac{a_{SICD}}{\rho_{cal}} \quad (2)$$

where aSICD is defined as the strength of the ICD, q is the volume flow rate of fluid mixture through the ICD at local conditions, which is equal to the volume flow rate through the ICD segment multiplied by a scaling factor that depends on the length of the device.
The density of the fluid mixture at local segment conditions is given in (3).

$$\rho_{mix} = \alpha_o \rho_o + \alpha_w \rho_w + \alpha_g \rho_g \quad (3)$$

where αo,w,g is the volume fraction of the free oil, water, gas phases at local conditions and ρo,w,g is the density of the oil, water, gas phases at local conditions (Schlumberger, 2013).
The viscosity of the fluid mixture at local segment conditions is given in (4)

$$\mu_{mix} = (\alpha_o + \alpha_w) \cdot \mu_{emul} + \alpha_g \cdot \mu_g \quad (4)$$

where μemul is the viscosity of the oil-water emulsion at local conditions and μg is the gas viscosity at local conditions. The calculation of μemul is described in "Emulsion viscosity" section (Schlumberger, 2013).

To include a series of these devices in a multi-segment well, the devices should be represented by segments branching off the tubing as shown in Figure 2. The grid block connections are located in the ICD segments instead of the segments representing the well tubing. The ICD segments should be given a very small length (of the order, say, of the wellbore radius). This length is not used in the pressure loss calculations, but it influences the location of the connections of the grid block in the reservoir. The ICD segments were given the same depth as their 'parent' tubing segments, so that there will be no hydrostatic head across them (Johnson and Oddie, 2004). The pressure loss across an ICD segment is reported as the friction pressure loss; the acceleration pressure loss is set to zero.

2.2 Emulsion Viscosity

The emulsion viscosity is a function of the local phase volume fractions in the segment and has differing functional forms at low water in liquid fractions (when oil is the continuous phase) and high water in liquid fractions (when water is the continuous phase) (Schlumberger, 2013). A critical water in liquid fraction as shown in figure 3 is used to select between (5) and (6).

$$\mu_{wio} = \mu_o \cdot \left(\frac{1}{1-\left(\frac{0.8415}{0.7480}\cdot\alpha_{wl}\right)}\right)^{2.5} \quad (5)$$

$$\mu_{oiw} = \mu_w \cdot \left(\frac{1}{1-\left(\frac{0.6019}{0.6410}\cdot\alpha_{ol}\right)}\right)^{2.5} \quad (6)$$

where μ_wio is the water-in-oil emulsion viscosity (when oil is the continuous phase), μ_oiw is the oil-in-water emulsion viscosity (when water is the continuous phase) and μ_o is the oil viscosity at local conditions. μ_w is the water viscosity at local conditions, α_wl is the local water in liquid fraction and α_ol is the local oil in liquid fraction.

The water-in-oil viscosity is subject to an upper limit expressed as a maximum ratio of water-in-oil viscosity to oil viscosity. This usually results in a 'plateau' region within which the water-in-oil viscosity is at its maximum permitted value as shown schematically in Figure 3, with the maximum viscosity ratio set at 5.0 (Schlumberger, 2013).
This upper limit also applies to the oil-in-water viscosity, but is less commonly encountered. At the critical water in liquid fraction there is a jump in emulsion viscosity as the continuous phase changes. Such a discontinuity would cause stability problems in the simulator and a transition region is defined about the critical water in liquid fraction to avoid this. In this region the emulsion viscosity is linearly interpolated between the water-in-oil and oil-in-water viscosities at the edges of the region; the viscosity is thus a continuous function of the water in liquid fraction. This transition region is presented schematically in Figure 3, with the linear interpolation shown in red between points B and C (Schlumberger, 2013).

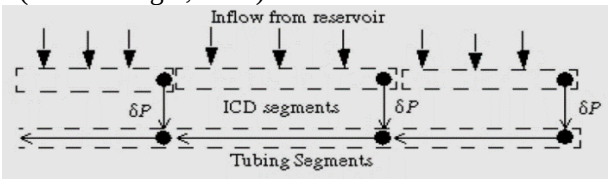


Figure 2. Segments ICDs along the well (Schlumberger, 2013).

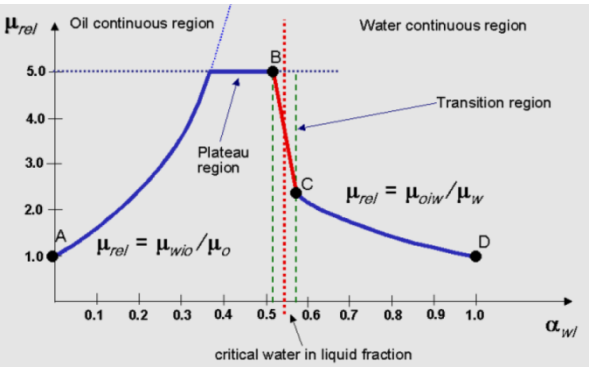


Figure 3. Phase Ttransition region about the critical water in liquid fraction (Schlumberger, 2013).

between the water-in-oil and oil-in-water viscosities at the edges of the region; the viscosity is thus a continuous function of the water in liquid fraction. This transition region is presented schematically in Figure 3, with the linear interpolation shown in red between points B and C (Schlumberger, 2013).

3 Case Study

A study was considered with reservoir conditions similar to the Troll field, Norway to illustrate the effect of ICD on oil recovery, reservoir sweep, delay in water breakthrough and decrease in water cut. Troll is a large subsea offshore Norway. The challenge is to drill and complete well in a way that gas and water do not have easy access to the production well (Henriksen et al., 2006). The main oil reservoir at Troll is the Late Jurassic Sognefjord Formation. This formation consists of Sandstone and siltstone with thickness of about 160m. The porosity vary between 30 -35% and permeability between 1 – 20D. The reservoir driving mechanism is mainly gas expansion and water drive. Horizontal wells are located close to the oil-water contact in order to reduce gas breakthrough (Henriksen et al., 2006).

Simulation was carried out for 3000 days. Water drive was achieved by connecting analytical aquifer (Fetkovich aquifer) at the bottom of the reservoir. Frictional pressure drop and variation in permeability will lead to non-uniform inflow profile along the production well (Aakre et al., 2013). ICDs are set at two segments along the production open hole section to distribute downhole pressure to optimize fluid inflow along the entire production interval. Water saturation profile shown in Figure 4 indicates that more water is produced at the 225m and 375m positions of the production well due to high permeability at these positions. To reduce water breakthrough, ICDs were placed at these positions. Each ICD joint is about 12m in length and about 3mm nozzle diameter. A base case without water ICD was considered for reference.

3.1 Geometry

Rectangular reservoir geometry was considered with the dimension 500m x 450m x 70m. The multi-segment horizontal production (PROD) well is of length 450m. The reservoir is heterogeneous with varying permeability from 1 to 20 Darcy. The areas of high permeability represents defeat in the reservoir as shown in figures 5 and 6.

3.2 Reservoir Conditions

The reservoir is heterogeneous and consists of water-wetted rock. Although the reservoir fluid consists of live black oil, gas production was not considered for simplicity. The composition of oil components is assumed to be constant relative to pressure and time. It is also assumed that the reservoir fluid is Newtonian and that Darcy's law applies. The reservoir conditions used for the simulation are summarized in Table 2.

3.3 Assumptions

The following assumptions were made regarding the inflow:

1. Darcy's law applies to the flow through the reservoir.
2. The flow into the well is at steady or pseudo-steady state.
3. The flow into the well is at steady or pseudo-steady state.
4. The distance between the well and the reservoir boundary is longer than the length of the well length.

The following assumptions were made about the ICDs:

1. There is no flow in the annulus parallel to the base pipe. This means that fluid flows from reservoir directly through ICD screens into the base pipe (Ouyang, 2009).
2. ICDs installed are of the same strength. This is the most common type of ICD application due to the relative simplicity of its design and installation operation (Henriksen et al, 2006). This is done in order to reduce the operational risks (Birchenko et al, 2010; Muradov et al, 2010).

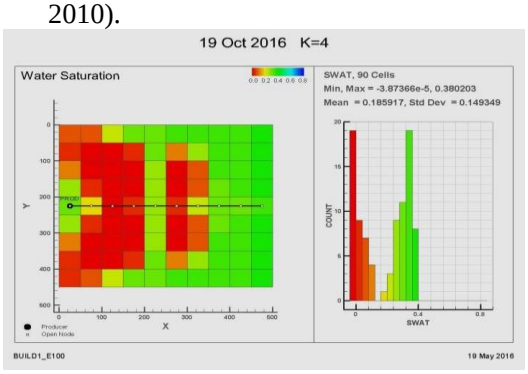


Figure 4. ICD positions along the well.

3.4 Initial Conditions

Initially, the reservoir is assumed to be in hydrostatic equilibrium consisting of only oil. The initial pressure is greater than the bubble point and water has much higher mobililty than oil. Table 1 shows the initial conditions considered during the simulation.

4 Result and Discussion

In this simulation, the effect of ICD completion on oil, water and gas production was investigated. Also the reservoir pressure trend recovery was discussed. A base case without ICD completion was considered as reference.

Table 1. Initial Conditions.

Initial condition	Value	Unit
Reservoir pressure	320	Bar
Bottomhole pressure	310	Bar
Bubble point pressure	182	Bar
Oil saturation	1	-
Water saturation	0	-
Gas saturation	0	-

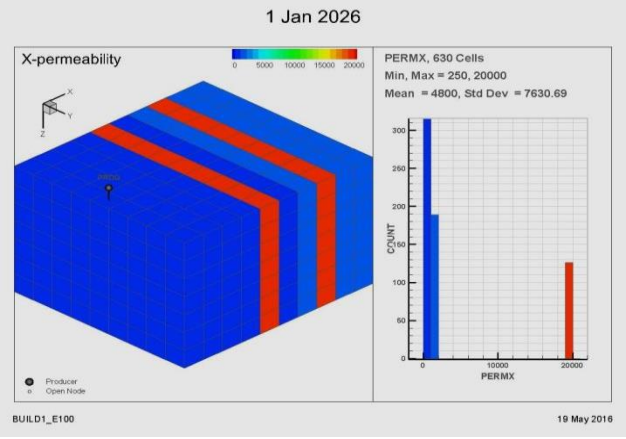


Figure 5. Reservoir geometry showing the distribution of X and Y permeability.

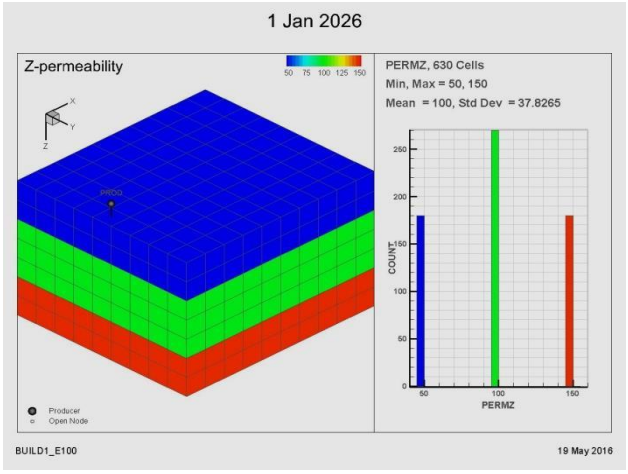


Figure 6. Reservoir geometry showing the distribution of Z- permeability.

4.1 Reservoir Pressure

Figure 7 shows the simulated reservoir pressure trend. The ratio of the total pressure drop without ICD completion to the total pressure drop with ICD is about 52. The high pressure drop for the case without ICD may be due to more reservoir depletion as a result of high water production. ICD tends to maintain the reservoir pressure by retaining water in the reservoir pore spaces.

4.2 Water Production

The water cut trend is shown in Figure 8. It is observed that water breakthrough is delayed for 262 days (about 66%) with the installation of ICD. Also the water cut is

Table 2. Reservoir Conditions.

Parameter	Value	Unit
Components	Oil, water, gas	-
Wettability	Water-wetted	-
Porosity	0.30	-
X Permeability	0.1 - 20	Darcy
Y Permeability	0.1 - 20	Darcy
Z Permeability	0.1-1	Darcy
Rock compressibility	5.0E-5@ 10Bar	/Bar
Oil gravity	35	°Api
Residual oil sat	0.3	-
Oil viscosity	10 @ 320Bar	cP
Water Density	1000	kg/m^3
Water viscosity	0.5	cP
Connate water sat	0.2	-
Gas density	1	kg/m^3
Well length	450	m
Target well flow rate	2000	Sm^3/day
ICD Length	12	m
ICD Strength	0.00021	$bar/(Rm^3/day)^2$
ICD nozzle diameter	3	mm
Simulation time	3000	days
No of Grids	630 (10x9x7)	-

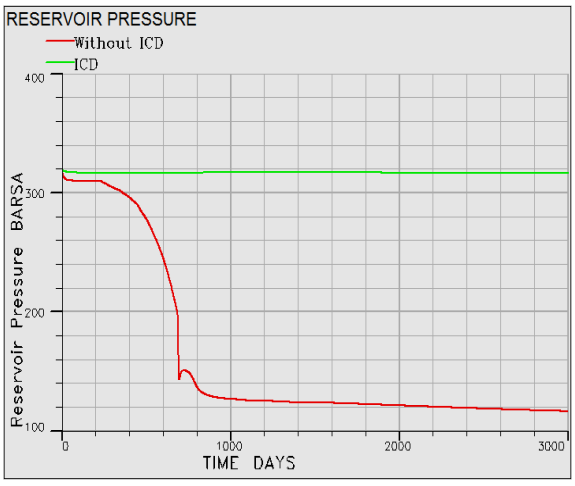


Figure 7. Reservoir Pressure Trend.

reduced with about 11% after 3000 days with the ICD completion. This would be attributed to the restriction imposed on water flow due to the additional pressure drop with the ICD.

4.3 Oil Production

Figure 9 shows the oil production rate with and without ICD respectively. Although the water breakthrough is delayed with ICD, the oil production rate is lower compared with the case without ICD. After water breakthrough, the production rate drops more rapidly for the case without ICD. This may be attributed to rapid water production as there is no restriction towards water production. Shock wave was propagated at about 690th day due to sudden opening of valve to match up the production target for the case without ICD. This shock wave can lead to very high pressure buildup which could make the system to fail. With the ICD, this phenomenon is annulled through its equalization effect on flow variation making the system stable throughout the production life.

Although well productivity is reduced by approximately 42%, there is an improved degree of inflow equalization through ICD completion. The accumulated oil production is shown in Figure 10. From the slope, production would be sustained more and the accumulated oil production expected to be higher over a long time with ICD completion.

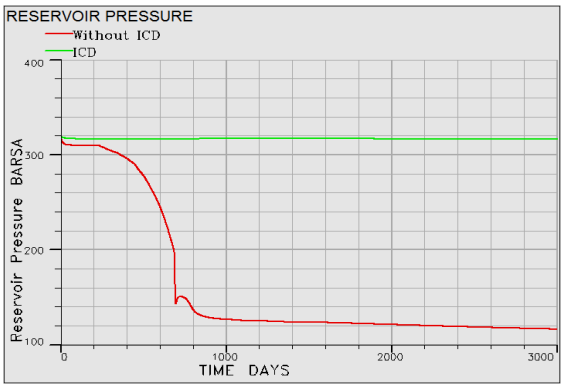


Figure 8. Reservoir Pressure Trend.

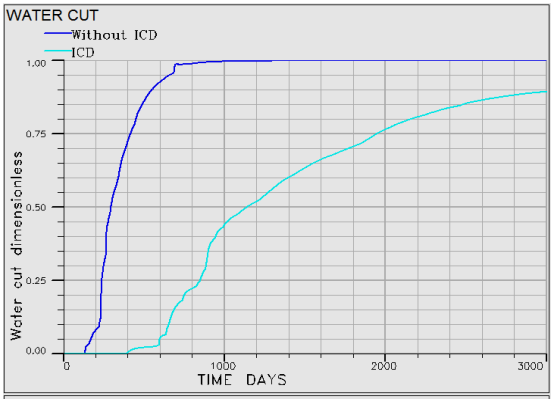


Figure 9. Trend of water cut.

4.4 Gas Production

Figure 11 shows the gas production rate with ICD and without ICD completions respectively. It can be seen that gas production rate is less with ICD completion throughout the production life. This may be attributed to rapid water production in the case without ICD as there is no restriction towards water production. Shock wave was propagated at about 690th day due to sudden opening of valve to match up production target for the case without ICD. This shock wave can lead to system failure as result of high pressure. This shock effect is not observed with ICD completion due to the restriction imposed by additional pressure drop and the equalization effect on flow variation. With ICD completion, the system is stable throughout the production life. There is about 51% decrease in gas production as depicted in Figure 12 with ICD completion. This increase in gas production for the case without ICD may reduce well performance and recovery significantly as oppose to ICD completion.

5 Conclusions

This paper presents the mathematical models used for the implementation of ICD in ECLIPSE reservoir simulator. A case study using similar reservoir conditions as Troll offshore Norway was simulated to illustrate the effect of ICD in a heterogeneous reservoir. Analysis of oil, water and gas production was made within a simulation period of 3000 days.

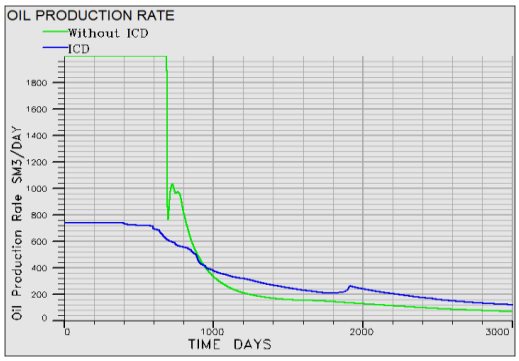


Figure 10. Trend of Oil Production Rate.

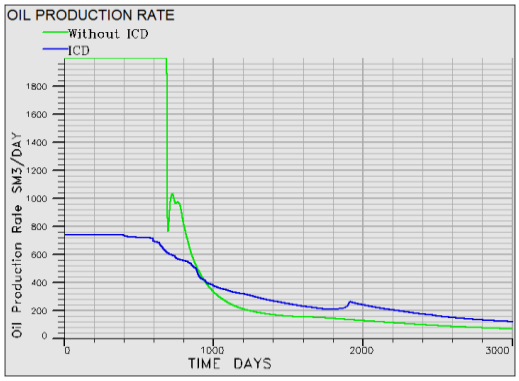


Figure 11. Trend of Oil Production Rate.

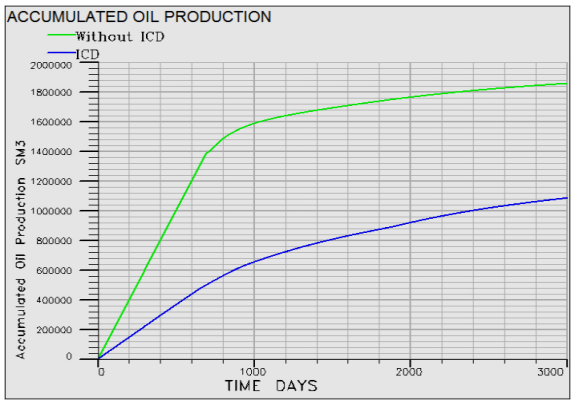


Figure 12. Trend of Accumulated Oil Production.

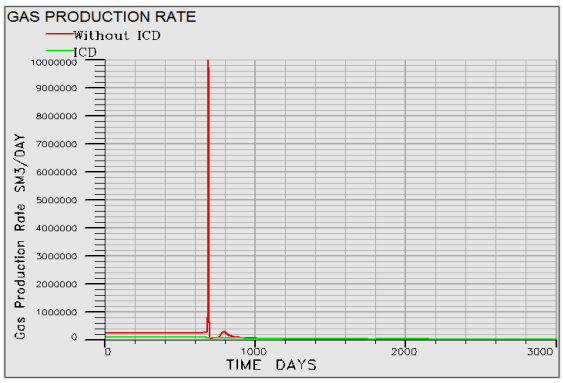


Figure 13. Trend of Gas Production Rate.

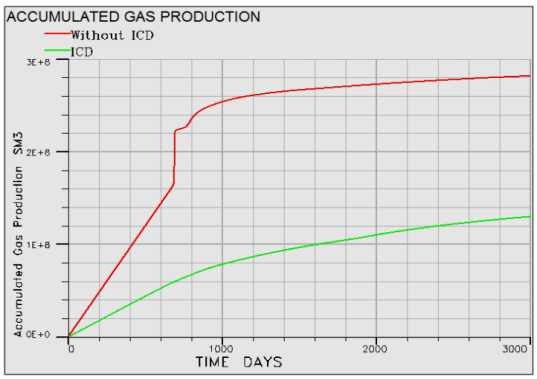


Figure 14. Trend of the accumulated gas production.

Result shows that with ICD completion, water breakthrough was delayed with 262 day and water cut after 3000 days was reduced by 11%. Despite the delay in water breakthrough, the oil production rate was reduced due to flow restriction by additional pressure drop with ICD completion. A trade-off between well productivity and inflow equalization is important. Although well productivity is reduced by approximately 42%, there is an improved degree of inflow equalization through ICD completion. Gas production was decreased by approximately 51% with ICD completion. With this reduction in gas production, well performance and ultimate recovery would improve. Result also indicates that the case with ICD completion sustains the reservoir pressure as water is forced to occupy the pore spaces of the reservoir.

It would be inferred that although ICD delays water and gas breakthrough, it could not stop the breakthrough. It would be appropriate to apply autonomous inflow control device instead, to stop gas and water breakthrough.

Acknowledgements

We would like to thank the management of University College of Southeast Norway and Inflow Control AS for providing the facility required for this study.

References

H. Aakre, B. Halvorsen, B. Werswick and V. Mathiesen. Smart well with autonomous inflow control valve technology. In *Proceedings of the SPE Middle East Oil and Gas Show and Conference*, 2013.

V. Birchenko, A. I. Bejan, A. Usnich and D. Davies. Application of inflow control devices to heterogeneous reservoirs. *Journal of Petroleum Science and Engineering*, 78(2): 534-541, 2011.

V. Birchenko, K. Muradov and D. Davies. Modeling of the Heel–toe Effect in a Horizontal Well with Inflow Control Devices. In *Proceedings of the ECMOR XII-12th European Conference on the Mathematics of Oil Recovery*, 2010.

V. Birchenko, K. Muradov and D. Davies. Reduction of the horizontal well's heel–toe effect with inflow control devices. *Journal of Petroleum Science and Engineering*, 75(1): 244-250, 2010.

J. Hallundbæk and P. Hazel. Inflow control in a production casing. *Google Patents*, 2016.

K. H. Henriksen, E. I. Gule and J. R. Augustine. Case study: the application of inflow control devices in the troll field. In *Proceedings of the SPE Europec/EAGE Annual Conference and Exhibition*, 2006.

C. D. Johnson and G. M. Oddie. Flow control regulation method and apparatus. *Google Patents*, 2004.

D. Krinis, D. E. Hembling, N. J. Al-Dawood, S. A. Al-Qatari, S. Simonian and G. Salerno. Optimizing Horizontal Well Performance In Non-Uniform Pressure Environments Using Passive Inflow Control Devices. In *Proceedings of the Offshore Technology Conference*, 2009.

V. Mathiesen, B. Werswick, H. Aakre and G. Elseth, G. Autonomous Valve- A Game Changer of Inflow Control in Horizontal Wells. In *Proceedings of the Offshore Europe*, 2011.

K. Muradov, V, Birchenko and D. Davies. Modeling of the Heel–toe Effect in a Horizontal Well with Inflow Control Devices. In *Proceedings of the 12th European Conference on the Mathematics of Oil Recovery*, 2010.

L. B. Ouyang. Practical consideration of an inflow-control device application for reducing water production. In *Proceedings of the SPE Annual Technical Conference and Exhibition*, 2009.

Schlumberger Limited. *ECLIPSE Reservoir Simulation Software - Technical Description*, 2013.

Domain-Specific Modelling of Micro Manufacturing Processes for the Design of Alternative Process Chains

Daniel Rippel[1] Michael Lütjen[1] Michael Freitag[1,2]

[1]BIBA – Bremer Institut für Produktion und Logistik GmbH at the University of Bremen, Germany, {rip,ltj}@biba.uni-bremen.de

[2]Faculty of Production Engineering, University of Bremen, Germany, fre@biba.uni-bremen.de

Abstract

In the context of an industrial production of micro components, the planning and configuration of process chains constitutes a major factor of success for the involved companies. Besides very small tolerances and high quality requirements, high production speeds have to be achieved. Moreover, so called size-effects introduce additional uncertainties to the planning process. While the modelling methodology "Micro - Process Planning and Analysis" provides a series of tools and methods to achieve a detailed planning and configuration of process chains in micro manufacturing, the high level of detail requires a comparably large amount of manual work, as well as a broad knowledge about available processes. Moreover, several processes can be substituted to achieve specific forms and shapes, while providing their own advantages and disadvantages for the overall production system. This article describes an extension to the methodology, which enables an automatic selection of suitable processes using geometry focused annotations. While these annotations only add minor efforts to the modelling process, they can be used to automatically derive alternative process chains. Particularly for production systems offering a broad range of processes, this extension reduces the manual effort in modelling and evaluating alternative process chains.

Keywords: micro manufacturing, process planning, process configuration, geometry oriented process chain design

1 Introduction

During the last years the demand for metallic micro parts has increased continuously. While they become increasingly smaller, their shape's complexity and level of functional integration constantly increases (Wulfberg *et al.*, 2010; Hansen *et al.*, 2006; Mounier, Bonnabel, 2013). Aside from more complex applications for micro components, an increasing application of these components within the growth markets of medical- and consumer-electronics constitutes a primary driver for this development (Mounier, Bonnabel, 2013). Besides the growing demand for Micro-Electro-Mechanical-Systems (MEMS), which are generally produced using methods from the semi-conductor industry, the demand for metallic micromechanical components increases similarly. These are generally used as connectors for MEMS, casings, or contacts. These micromechanical components are usually manufactured by applying processes from the areas of micro forming, micro injection, micro milling etc. (Hansen *et al.*, 2006; Fu, Chan, 2012). Particularly, cold forming processes constitute a performant option for the realization of an economic mass production of metallic micromechanical components. These types of processes generally provide high throughput rates at comparably low energy and waste costs (DeGarmo *et al.*, 2003).

An industrial production of such components is usually characterized by high throughput rates up to several hundred parts per minute (Flosky, Vollertsen, 2014), whereby very small tolerances have to be achieved. These tolerances result from the components' small dimensions, which are by definition smaller than one millimeter in at least two geometrical dimensions (Geiger *et al.*, 2001). Moreover, so-called size-effects can result in increasing uncertainties and unexpected process behaviors when processes, originating from the macro domain, are scaled down to the micro level (Vollertsen, 2008). Additionally, micro manufacturing is an active and relatively young field of research for scientists as well as industrials, leading to a continuous development of new or enhanced processes and machines.

As a result, the planning and configuration of process chains constitutes one major success factor for an industrial production of metallic micromechanical components (Afazov, 2012). To cope with the occurrence of size-effects, companies require a highly precise planning not only of the single processes configurations, but also spanning the complete process chain. Thereby, interrelationships between processes, materials, tools and devices have to be considered. Small variations in single parameters can have significant influences along the process chain and can

finally impede the compliance with the respective tolerances (Rippel *et al.*, 2014).

This article describes an extension to the Micro – Processes Planning and Analysis (µ-ProPlAn) methodology, which is designed to provide the necessary tools and procedures for an accurate planning and configuration of process chains within the micro domain (see e.g. (Rippel *et al.*, 2014; Rippel *et al.*, 2014b)). The methodology itself provides a set of methods to graphically model, plan and evaluate process chains across different levels of granularity. Nevertheless, due to the emergence of new technologies, there exist several possible processes to achieve the same product or work piece feature. For example, a hole can be placed within a metal sheet via drilling, laser-chemical or electrical ablation. The manual creation and evaluation of alternative process chain models can be time consuming and requires a broad knowledge of available technologies as well as of their different characteristics. In order to facilitate the search for the most economical solution, this article describes an extension to µ-ProPlAn that allows an automatic selection of viable alternatives based on geometrical features of the desired work piece. The remainder of this section provides a short description of size-effects, followed by an overview of the state of the art in process planning within the micro domain. The next section gives a short introduction to the µ-ProPlAn methodology and its components. Afterwards, the article describes the extension, its meta-model and a short example for its application. The article closes with a discussion of the extension as well as a description of planned future work on this topic.

1.1 Size-Effects in Micro Cold Forming

While cold forming processes are well established in mass production within macro manufacturing, scaling these processes down to the micro domain is only possible to a certain degree. With a decreasing scale of the machines, tools and work pieces, so called size-effects begin to emerge, which require changes and adaptations to the processes.

Vollertsen defines size effects as “deviations from intensive or proportional extrapolated extensive values of a process, which occur when scaling the geometrical dimensions” (Vollertsen, 2008). In this context, he defines intensive values as parameters, which are not expected to change due to a change of an object’s mass (e.g. its temperature or its density) as well as extensive values that are expected to vary (e.g. the object’s inertia force or its heat content). In general, size effects occur due to the inability to scale all relevant parameters equally (Vollertsen, 2008). For example, the downscaling of a metal sheet can result in stronger variations of its density due to local defects, although the density is considered an intensive variable. In macro manufacturing these variations can be ignored, while they can have drastic influences in micro manufacturing. Moreover, technical limitations can further facilitate the occurrence of size effects. For instance, the downscaling of mechanical grippers is limited by technical factors and only possible to a certain degree. For tiny work pieces, Van-der-Waals forces between the gripper and the work piece will eventually overcome the gravitational force at a certain point of miniaturization. As a result, the gripper will not be able to release the work piece without aid. Vollertsen defines three distinct categories of size effects (Vollertsen, 2008):

- Density size-effects occur, when the density of a material is held constant, while scaling down its geometrical dimensions. For instance, local defects become more serious with a continuing miniaturization. Thereby, the distribution of local defects within a material can lead to more delimited sets of good and bad parts.
- Shape size-effects occur due to the increasing ratio of an object’s total surface area, compared to its volume. An example of this category is provided by the described imbalance of the adhesive force in relation to the gravitational force.
- Micro structure size-effects occur because micro structural features (e.g. the grain size or the surface roughness) cannot be scaled down the same way as the geometrical size of an object.

The occurrence of size-effects requires precise planning and configuration of all relevant technical parameters throughout a process chain. Due to size-effects and the continuous development of new processes and technologies for micro manufacturing, interrelations between those parameters can rarely be described comprehensively or are entirely unknown in several cases.

1.2 Process Planning and Configuration in Micro Manufacturing

Existing literature, describes only very few approaches that enable a joint planning of process chains as well as the technological and logistic configuration of the involved processes. During the last years, different articles focused on the configuration of specific processes (compare e.g. (Afazov, 2013)). Thereby, these approaches rely on detailed studies of the corresponding processes and are usually supported by highly detailed physical models in form of finite element simulations (e.g. (Afazov, 2012; Pietrzyk *et al.*, 2008)). A different type of approach found in the literature focusses on the use of sample data (historical or experimental) as templates for the configuration of processes (e.g. (Sabotin *et al.*, 2009)). Although both of these approaches allow for a precise configuration of single processes, the interrelations between different processes within a process chain cannot be considered easily.

Moreover, the construction of finite element simulations as well as the direct application of historical information requires a comprehensive understanding of the processes and of the physical backgrounds, which in many cases is unavailable to the process planner due to size-effects or the novelty of processes.

In general, methods such as event-driven process chains, UML or simple flow charts are used in the context of process chain planning. While these methods do not enable a configuration of processes, Denkena et.al. proposed an approach that indirectly addresses this topic (Denkena *et al*., 2006). This approach relies on the modelling concept for process chains (Denkena, Tönshoff, 2011). In this concept, a process chain consists of different process elements, again consisting of operations. These operations are interconnected by so-called technological interfaces, which generally describe sets of pre- or post-conditions for each operation. While these interfaces can be configured manually, Denkena *et al*. extended this concept by proposing the use of physical, numerical or empirical models to estimate the relationships between the pre- and post-conditions (Denkena *et al*., 2006; Denkena *et al*., 2014). Although this approach enables configuration of the processes, the creation of these models requires a very detailed insight into the processes as stated before. Moreover, within the micro domain, a single model can easily be unsuitable to capture all relevant interrelations between process-, machine-, tool- and work piece-parameters, particularly under the influence of size-effects.

Based on the literature review, it can be concluded, that there is no method enabling a joint planning and configuration of process chains for the micro domain, which provides the necessary level of detail and generality to cope with size-effects as well as with the continuous development of new technologies in this domain.

2 µ-ProPlAn

The modelling methodology µ-ProPlAn covers all phases from the process and material flow planning to the configuration and evaluation of the processes and process chain models (Rippel *et al*., 2014). It enables an integrated planning of manufacturing, handling and quality inspection activities at different levels of detail, starting on the level of process chains, down to the level of cause-effect relations between single parameters. The methodology consists of a modelling concept, an accompanying procedure model as well as different methods for the evaluation of the modelled production systems' technological feasibility and logistic performance. All of these aspects are integrated in a software prototype.

The modelling concept consists of three views, which represent different levels of detail. The first view focusses on the top-level process chains, whereby the second view extends these process models by adding information on the specific material flow. The third view is used to further detail all relevant material flow elements (e.g. machines or work pieces) in terms of their parameters and the parameters' interrelations (cause-effect networks). Figure 1. depicts the most important entities used within the µ-ProPlAn models and their hierarchical composition.

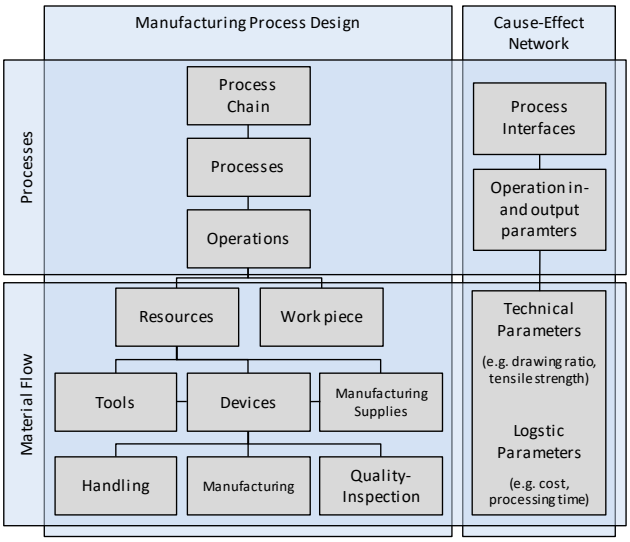
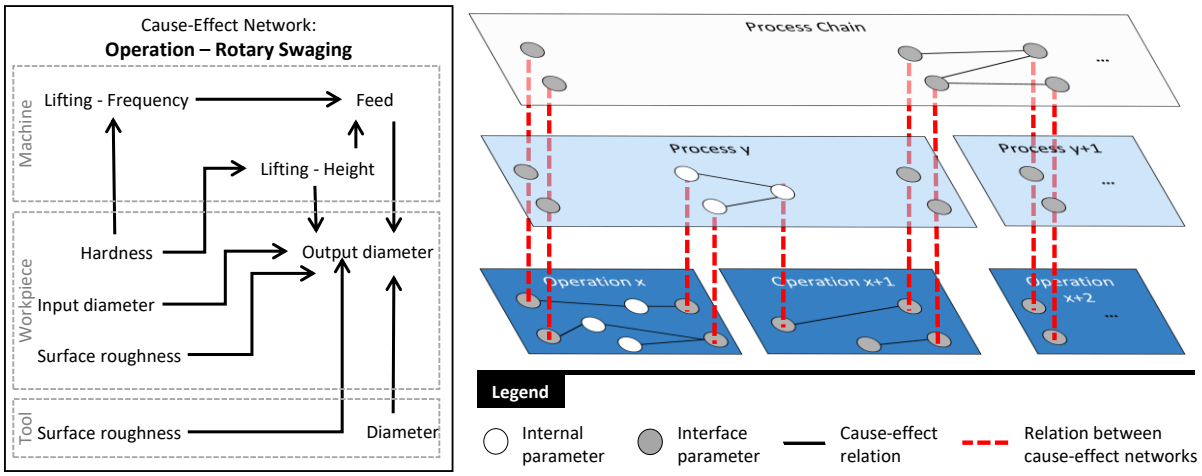


Figure 1. Major Entitties used within µ-ProPlAn's models (Rippel *et al*., 2014).

The process view applies a notation closely conforming to the widely used notation of process chains, as described in the state of the art (Denkena, Tönshoff, 2011). In difference to the classical approach, process elements as well as operations are connected using extended process interfaces. These interfaces also include logistic parameters, asides from the technical parameters used within the classic approach. Additionally, operations act as interfaces to the next view: the material flow view.

On the one hand, the material flow view allows modeling of all material flow objects present in the modelled production scenario (i.e. machines/devices, work pieces, tools, operating supplies or workers). On the other hand, it is used to further detail operations by assigning those material flow objects that are used to conduct the operation. This enables modelling of specific production scenarios, for example, factories or production sites with specified resources as well as specific routes for each product (process chain). As a result, it becomes possible to evaluate the models using discrete-event material flow simulations regarding logistic aspects. For example, in case of an existing production system, µ-ProPlAn can be used to assess the impact of new process chains (e.g. new products) on existing production plans or on the performance of the production system in general. Therefore, the µ-ProPlAn software prototype can directly transform these models



a) (b)

Figure 2. Hierarchy of cause-effect networks: (a) Exemplary composition of an operation from the networks of the corresponding machine, tool and work piece. (Rippel *et al*., 2014b); (b) Composition of higher level cause-effect-networks for process elements and process chains (Rippel *et al*., 2014).

into discrete-event simulation models for the *jasima*[1] library and execute them. This allows an early detection of bottleneck machines or other undesired effects early during the planning stages. Consequently, this supports the selection of suitable resources, machines or devices for the new process chain.

The third view focusses on the configuration of the processes and process chains using cause-effect networks to describe the interrelationships between relevant process parameters. In contrast to holistic approaches like those described in the state of the art, each network consists of a set of parameters and a set of cause-effect relationships, forming a directed graph. The set of parameters consists of all technical and logistic characteristics that are relevant to describe the respective object's influence on the production process. In case of work pieces these are e.g. costs per unit, material properties or geometrical characteristics. For machines, these parameters include velocities, forces or other characteristics that can be set, calculated or measured (compare Figure 2.a for an example). From a modelling perspective, the cause-effect networks are modelled hierarchically. Each material flow object (work pieces, machines, tools, workers, etc.) holds its own cause-effect network or at least a set of describing parameters. When combining these single elements to operations, process elements or process chains, higher level cause-effect networks are created by introducing additional relationships between parameters of the networks or by connecting them through previously specified process interfaces (compare Figure 2.b).

The creation of cause-effect networks is divided into two steps: the qualitative modelling and the quantification. The quantitative model of the corresponding network is created by collecting all relevant parameters and denoting their influences among each other (cf. Figure 2.a). The second step concerns the quantification of the cause-effect networks. The objective is to enable the propagation of different parametrizations throughout the network. For example, this propagation allows estimating the process' outcome for different materials or machining strategies. In Figure 2.a, the use of a different material for the forging tool will result in a different surface roughness, thus influencing the overall process' output diameter. Through quantifying the cause-effect relationships, it is possible to estimate the results of parameter changes to all connected parameters along complete process chains.

In case of simple or well-known relations, µ-ProPlAn allows to input mathematical formulas directly. Each formula thereby calculates a parameter's value based on its input parameters' values. For example, a continuous manufacturing process will have a duration according to the total length of the work piece divided by the selected feed velocity. More complex but well-established cause-effect relations could be included from literature, e.g. the calculation of the static friction of a tool and work piece based on their surface roughness. Nevertheless, in the area of micro manufacturing, different parameters can have a more significant impact than in the macro domain, resulting in the inclusion of parameters that can be neglected in macro manufacturing. In addition, size effects may induce a different behavior than usually observed. Therefore, it is often impossible to comprehensively describe all parameters and cause-effects relations directly. As a result, µ-ProPlAn proposes the derivation of prediction models from experimental- or from production data. Therefore, the

[1] JAva SImulator for MAnufacturing and logistics: https://code.google.com/archive/p/jasima/

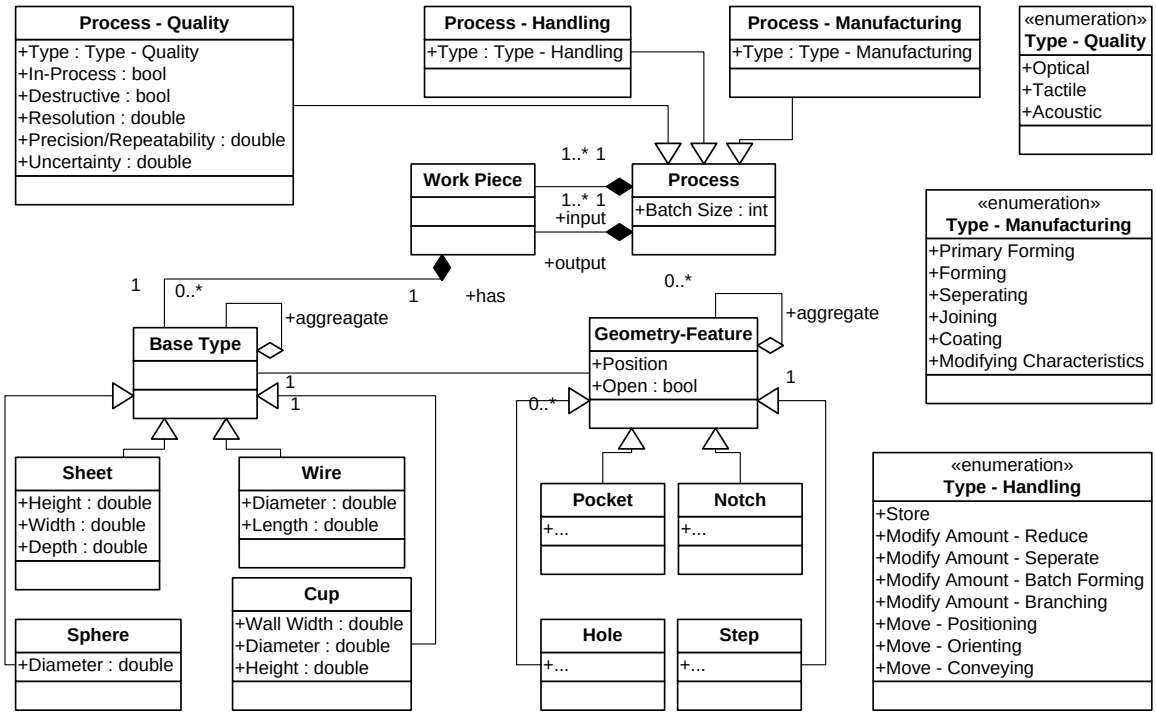


Figure 3. Simplified extension to the µ-ProPlAn meta model using the UML notation. Attributes denote additional parameters for the respective cause-effect networks.

software prototype offers a variety of regression techniques from the area of Artificial Intelligence (e.g. Support-Vector Machines or Artificial Neural Networks) as well as statistical methods (e.g. the deduction of least square linear or polynomial regression models or locally weighted linear regression models). In practice, the application of locally weighted linear regression models has shown high accuracy in estimating unknown or hard-to-describe cause-effect relations (see e.g. (Rippel *et al.*, 2014; Rippel *et al.*, 2014b)). After quantification, each parameter contains a model (mathematical or prediction) that allows to estimate its value based on the remaining values in the cause-effect network.

As a result, these networks enable an in-depth evaluation of different process configurations (e.g. the use of different materials or different production velocities) and to assess the impacts of different choices on follow-up processes or the production system in general. As cause-effect networks and material flow elements are closely connected, µ-ProPlAn can directly reflect changes to the configuration within the material flow simulation and evaluate these configurations e.g. regarding work-in-progress-levels, lead times or the products' estimated qualities.

While the methodology itself enables a highly precise planning of the process chains, as well as the configuration and evaluation of these chains, the high precision comes at the cost high modelling efforts. While the initial creation of material flow objects and their respective cause-effect networks can be conducted prior to the actual planning of a process chain, the introduction of new process chains (i.e. products) to an existing production system still requires a high amount of manual work. Thereby the production planner needs to select those processes and machines that result in the most economic overall production in terms of product quality as well manufacturing times and costs. In micro manufacturing, there exist a growing range of processes that can achieve similar results. As machines are often comparably small, production sites often house several different machines or processes, which leads to a possibly broad pool of alternative process chains. Nevertheless, each process provides its own characteristics in terms of velocity, cost, or other properties that can have a positive or a negative influence on the overall production system or on the products manufacturing process and characteristics. As the manual creation of alternative process chains can be quite time consuming and require a broad knowledge of available processes as well as their advantages and disadvantages, the next section describes an extension of the µ-ProPlAn methodology that can be applied for an automatic selection and creation of process chain alternatives based on geometrical features of the desired product.

3 Geometry oriented development of process chains

The geometry oriented process chain design provides an alternative approach of modelling and designing process chains in µ-ProPlAn. The current methodology assumes a manual modelling and evaluation of each alternative

chain. The geometry-oriented approach focusses on additional annotations to the modelled process, in order to select and combine suitable processes. Thereby, each process is annotated with information on its capabilities and limitations. In addition, work pieces can be specified by their geometrical features. Using a constraint based search, suitable processes can be selected, combined and evaluated automatically. For these annotations, additional parameters are introduced to the cause-effect networks in order to describe which geometries can be achieved by a process to which extend as well as the respective pre-conditions. Basically, the µ-ProPlAn meta model is adapted twofold: First, processes need to be annotated with their type (what they do). Second, each process requires annotation which geometrical features can be achieved under which circumstances.

The type of a process describes its function according to the standards DIN 8580 and VDI 2860. For manufacturing processes the type describes if it is a (primary) forming process, a separating or joining process, if it is a coating process or if it just modifies the work pieces characteristics without changing the geometry. For handling processes the type determines e.g. if it is a positioning or conveying process. Besides the definition of the type (optical, tactile, acoustic), quality inspection processes additionally require information if the process is destructive, if it can be used in-process as well as its resolution and uncertainties.

The second adaptation describes each process' pre- and post-conditions with respect to geometrical features. Thereby, each process gets annotated with a set of work pieces that can serve as input to the process as well as a set of work pieces which result from the process application to an input. In extension to the current meta model, each work piece has a base geometry, describing e.g. if the work piece is a sheet, a sphere, a wire etc. Figure 3. lists some examples of base geometries with their respective parameters in the lower left part. This list can be extended for a given application or model in order to include additional base geometries as needed. Each base geometry can be combined with additional base geometries to compose more complex work pieces. Figure 4.a for example shows a combination of a wire with a cone shaped base geometry to represent a work piece that could act as a plunger for a micro valve. Each base geometry can be assigned zero or more geometrical features such as holes, steps, pockets or notches. Figure 4. lists some examples derived from the STEP-NC standard. While this standard focusses on the feature-based description of machining strategies, it offers comprehensive descriptions of geometrical features as well as their characteristics (parameters). To keep the image simple, these parameters have been omitted in Figure 4. as they can be directly taken from the STEP-NC standard. Each of these features can again be assigned zero or more features on their own in a hierarchical manner. For example, this allows modelling of a block (Sheet) that has a pocket on its top side (Pocket), which again contains a hole (Hole) in its center.

By using these annotations, it becomes possible to apply constraint based search algorithms to select suitable processes. During the modelling process, the designer specifies the hierarchical order of base geometries and geometrical features required for the final product. The software tool can search through all stored processes within the model and select those, which can achieve each feature. In the example in Figure 4.b there is only one alternative to join a sphere to a wire (melting) but there exist two alternatives to form this sphere into a cone. As the characteristics of each base geometry and each feature are expressed as parameters integrated into the processes, cause-effect networks, the

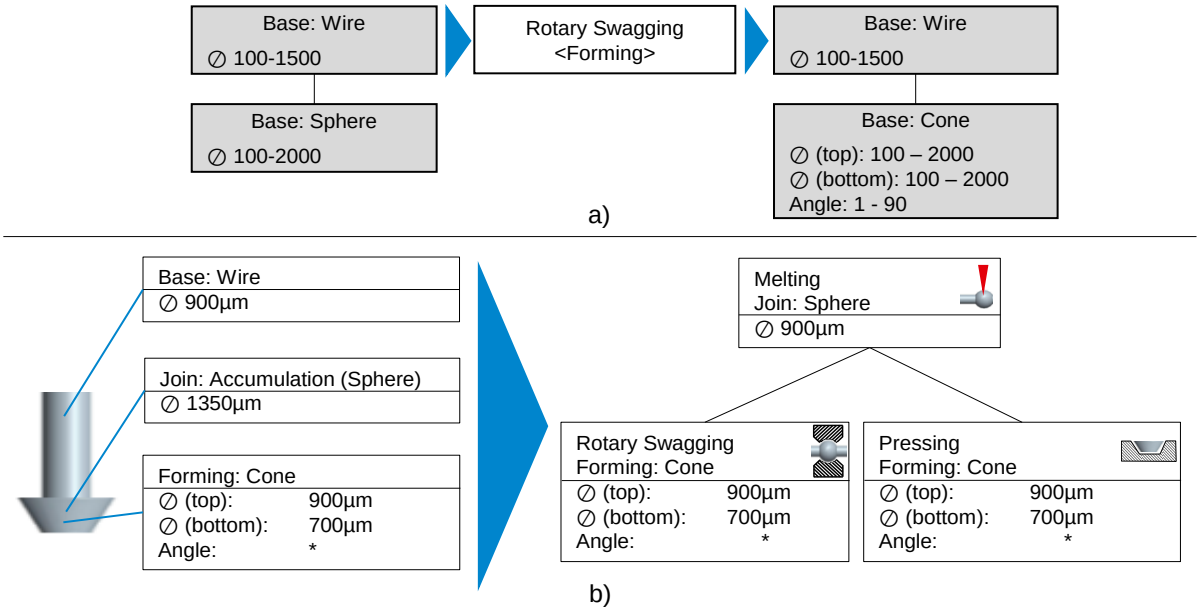


Figure 4.(a) Examples for the geometry based annotations (b) Example for the automatic selection of processes

configuration of each alternative process chain can be performed the usual way. As a result, technical characteristics (e.g. material or surface properties) can be estimated through the cause-effect networks, while logistic properties of different alternatives can be estimated by transforming and simulating the models using µ-ProPlAns discrete-event material flow simulation.

4 Discussion and Future Work

Concerning the design and configuration of process chains in micro manufacturing, the selection of suitable machining processes and devices constitutes a major influential factor on the production's performance. Following the trends of desktop manufacturing as well as of the continuous development of new or improved processes for micro manufacturing, a broad range of alternate processes can be available. As a result, the manual creation and evaluation of alternative process chains can easily become a time consuming task, prone to errors. This article proposes an adaptation to the µ-ProPlAn modelling methodology, focusing on the annotation of modelled processes, in order to enable an automatic derivation of alternative process chains, as well as their evaluation. This extension uses parts of commonly known standards like STEP-NC or different DIN and ISO standards to structure the required information and seamlessly integrate it with the cause-effect models used within µ-ProPlAn.

The changes to the µ-ProPlAns meta model provide a suitable approach to structure all the information required for an automatic selection of processes. As most of this information was already required in order to perform a configuration of the process chains in µ-ProPlAn's current modelling approach (e.g. information which dimensions should be achieved with respect to the work pieces form), the annotation of the processes and work pieces only induces minor additional efforts. Moreover, the changes provide a better way to structure this information, which facilitates the models' consistency and thus renders them easier to maintain and extend. While the possibility to search for alternative processes is restricted by the number of stored processes, the use of additional databases could allow the use of process templates across several models. Future work will focus on the conceptualization and implementation of such a database as well as on the identification of a more general list of base geometries for micro manufacturing.

Acknowledgements

The authors gratefully acknowledge the financial support by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for the Subproject C4 'Simultaneous Engineering' within the CRC 747 (Collaborative Research Center) "Mikrokaltumformen - Prozesse, Charakterisierung, Optimierung".

References

S. M. Afazov, A. A. Becker, and T. H. Hyde. Development of a Finite Element Data Exchange System for chain simulation of manufacturing processes. In *Advances in Engineering Software,*47(1): 104 - 113, 2012.

S. M. Afazov. Modelling and simulation of manufacturing process chains. In *CIRP Journal of Manufacturing Science and Technology*, 6(1): 70 - 77, 2013.

E. P. DeGarmo, J. T. Black, and R. A. Kohser. *Materials and Processes in Manufacturing*, 9th ed.: Wiley, 2003.

B. Denkena, H. Rudzio, and A. Brandes. Methodology for Dimensioning Technological Interfaces of Manufacturing Process Chains. In *CIRP Annals - Manufacturing Technology*, 55(1): 497 - 500, 2006.

B. Denkena and H. K. Tönshoff. Prozessauslegung und -integration in die Prozesskette. In *Spanen - Grundlagen*, B. Denkena and H. K. Tönshoff, Eds., Springer Verlag, Berlin, Heidelberg, pp. 339 - 362. 2011

B. Denkena, J. Schmidt, and M. Krüger. Data Mining Approach for Knowledge-based Process Planning. In *Procedia Technology* 15(1): 406 - 415, 2014.

H. Flosky and F. Vollertsen. Wear behaviour in a combined micro blanking and deep drawing process. In *CIRP Annals - Manufacturing Technology* 63(1): 281 - 284, 2014.

M. W. Fu and W. L. Chan. A review on the state-of-the-art microforming technologies. In *The International Journal of Advanced Manufacturing Technology,* 67(9): 2411 - 2437, 2012.

M. Geiger, M. Kleine, R. Eckstein, N. Tieslerl, and U. Engel. Microforming. In *CIRP Annals - Manufacturing Technology,* 50(2): 445-462, 2001.

H. N. Hansen, K. Carneiro, H. Haitjema, and L. De Chiffre. Dimensional Micro and Nano Technology. In *Annals of the CIRP Annals*, 55(2): 721-734, 2006.

E. Mounier and A. Bonnabel. Press Release Emerging MEMS, 29. August 2013. Yole Development, 2013.

M. Pietrzyk, L. Madej, and S. Weglarczyk. Tool for optimal design of manufacturing chain based on metal forming. In *CIRP Annals - Manufacturing Technology,* 57(1): 309 - 312, 2008.

D. Rippel, M. Lütjen, and B. Scholz-Reiter. A Framework for the Quality-Oriented Design of Micro Manufacturing Process Chains. In *Journal of Manufacturing Technology Management,* 25(7): 1028 - 1048, 2014.

D. Rippel, E. Moumi, M. Lütjen, B. Scholz-Reiter, and B. Kuhfuß. Application of Stochastic Regression for the Configuration of a Micro Rotary Swaging Process. In *Mathematical Problems in Engineering,* 2014b, http://dx.doi.org/10.1155/2014/360862.

I. Sabotin, J. Valentincic, M. Junkar, and A. Sluga. Process planning system for micro-products. In *Proceedings of the 10th International Conference on Management of Innovative Technologies*, p. 8, 2009.

F. Vollertsen. Categories of size effects. In *Production Engineering,* 2(4): 377-383, 2008.

J. P. Wulfsberg, T. Redlich, and P. Kohrs. Square Foot Manufacturing: a new production concept for micro manufacturing. In *Production Engineering - Research and Development,* 4(1): 75 - 83, 2010.

API for Accessing OpenModelica Models from Python

B. Lie[1] S. Bajracharya[2] A. Mengist[2] L. Buffoni[2] A. Kumar[2] M. Sjölund[2] A. Asghar[2] A. Pop[2] P. Fritzson[2]

[1]University College of Southeast Norway, Porsgrunn, Norway {Bernt.Lie}@hit.no
[2]Linköping University, Sweden {peter.fritzson}@liu.se

Abstract

This paper describes a new API for operating on Modelica models in Python, through OpenModelica. Modelica is an object oriented, acausal language for describing dynamic models in the form of Differential Algebraic Equations. Modelica and various implementations such as OpenModelica have limited support for model analysis, and it is of interest to integrate Modelica code with scripting languages such as Python, which facilitate the needed analysis possibilities. The API is based on a new class *ModelicaSystem* within package OMPython of OpenModelica, with methods that operate on instantiated models. Emphasis has been put on specification of a systematic structure for the various methods of the class. A simple case study involving a water tank is used to illustrate the basic ideas.

Keywords: OpenModelica, Modelica, Python, PythonAPI

1 Introduction

Modelica is a modern, equation based, acausal language for encoding models of dynamic systems in the form of differential algebraic equations (DAEs), see e.g. (Fritzson, 2014) on Modelica and e.g. (Brenan et al., 1987) on DAEs. *OpenModelica*[1] (Fritzson et al., 2006) is a mature, freely available toolset that includes *OpenModelica Connection Editor* (flow sheeting, textual editor with debugging facilities, and simulation environment) and the *OMShell* (command line execution, script based execution). OpenModelica Shell supports commands for simulation of Modelica models, for use of the Modelica extension *Optimica*, for carrying out analytic linearization via the Modelica package *Modelica_LinearSystem2*, and for converting Modelica models into Functional Mock-Up Units (FMUs) as well as for converting FMUs back to Modelica models. A tool *OMPython* has been developed and communicates with OpenModelica via CORBA, (Ganeson, 2012; Ganeson et al., 2012). Essentially, OMPython is a Python package which makes it possible to pass OpenModelica Shell commands as strings to a Python function, and then receive the results back into Python. This possibility does, however, require good knowledge of OpenModelica Shell commands and syntax. A tool, *PySimulator*,[2] has been developed to ease the use of Modelica from Python, (Pfeiffer et al., 2012; Ganeson et al., 2012). Essentially, PySimulator provides a GUI based on Python, where Modelica models can be run and results can presented. It is also possible to analyze the results using various packages in Python, e.g. FFT analysis. However, PySimulator currently does not give the user full freedom to integrate Modelica models with Python and use the full available set of packages in Python, or freely develop one's own analysis routines in Python.

Modelica and OpenModelica Shell in themselves have relatively little support for advanced analysis of models. Examples of such desirable analysis capabilities could be (i) study of model sensitivity, (ii) random number generation and statistical analysis, (iii) Monte Carlo simulation, (iv) advanced plotting capabilities, (v) general optimization capabilities, (vi) linear analysis and control synthesis, etc. Scripting languages such as MATLAB and Python hold most of these desirable analysis capabilities, and it is of interest to integrate Modelica models with such script languages. The free *JModelica.org* tool includes a Python package for converting Modelica models to FMUs, and then for importing the FMU as a Python object. This way, Modelica models can essentially be simulated from Python — Optimica is also supported. It is possible to do more advanced analysis with JModelica.org[3] via *CasADi*, see e.g. (Perera et al., 2015a,b). However, the possibilities in the work of Perera et al. use an old version of JModelica.org. It would be more ideal if these possibilities were supported by the tool developer.

It is thus of interest to develop an extension of OMPython which enables simulation and analysis of Modelica models with a better integration with the Python language, and in particular that such an extension is provided by the OpenModelica developers. A Python API[4] for controlling Modelica simulation and analysis from Python was proposed in February 2015[5]. Based on this proposal, a first version of a Python API has been implemented (Bajracharya, 2016), and has then been further revised. This paper discusses the API, and illustrates how it can be used for automatic analysis of Modelica models from Python, exemplified by a simple water tank

[1] www.openmodelica.org
[2] https://pypi.python.org/pypi/PySimulator
[3] www.JModelica.org
[4] API = Application Programming Interface
[5] *Python API for Accessing OpenModelica Models*, by B. Lie, February 20, 2015, communicated to P. Fritzson at Linköping University.

model. The paper is organized as follows. In Section 2, an overview of the API is given. In Section 3, use of the API is illustrated through simple analysis of a nonlinear reactor model. In Section 4, the API is discussed, some conclusions are drawn, and future work is discussed. Appendices hold details of the nonlinear reactor model.

2 Overview of Python API

2.1 Goal

Modeling and the use of Modelica with Python is of interest to a wide range of engineering disciplines. The computer science threshold of using Modelica with Python should be low. Ideally, the OMPython extension should work with simple one-click Python installations such as Anaconda[6] and Canopy[7]. Furthermore, the extension should support both 32 bit and 64 bit OpenModelica, work with both 32 bit and 64 bit Python, with Python 2.7 and Python 3.X, and on platforms Windows, OSX and Linux. These requirements e.g. imply that results should be returned as standard Python structures. However, it is reasonable that the OMPython extension depends on the NumPy package. Because Python has excellent plotting capabilities e.g. via Matplotlib, the OpenModelica Shell facility for plotting results should not be implemented — this is more naturally handled directly in Python.

2.2 Installing the OMPython Extension

Under Windows, the new OMPython extension will be automatically installed in a file `__init__.py` in directory `share\omc\scripts\PythonInterface\OMPython` in the OpenModelica directory when OpenModelica is downloaded and installed. In order to activate the extension, the user must next run the command `python setup.py install` from the command line in the directory of the `setup.py` file, which is in the `PythonInterface` subdirectory. It follows that in order to activate the extension, the user must first install Python on the relevant computer. Under Linux/OSX, OMPython is part of `pip` (`pypi`) and is not shipped with the OpenModelica installer.

2.3 Status

Currently, the Python API is in a development status and has been tested with 32 bit Python 2.7 from the Anaconda installation in tandem with 32 bit OpenModelica v. 1.9.4 under Windows 8.1 and OpenModelica v. 1.9.6 under Windows 10, and a modified `__init__.py` file. OpenModelica uses CORBA for communication, and CORBA compatibility needs some refinement. The code is somewhat unstable when run from the Spyder IDE used with the Anaconda installation, but runs fine from Jupyter notebooks.

[6] `www.continuum.io/downloads`

[7] `www.enthought.com/products/canopy`

2.4 Description of the API

The API is described in the subsections below.

2.4.1 Python Class and Constructor

The name of the Python *class* which is used for operation on Modelica models, is *ModelicaSystem*. This class is equipped with an object constructor of the same name as the class. In addition, the class is equipped with a number of methods for manipulating the instantiated objects.

In this subsection, we discuss how to import the class, and how to use the constructor to instantiate an object.

The object is imported from package *OMPython*, i.e. with Python commands[8]:

```
>>> from OMPython import ModelicaSystem
```

Other Python packages to be used such as `numpy`, `matplotlib`, `pandas`, etc. must be imported in a similar manner.

The object constructor requires a minimum of 2 input arguments which are strings, and may need a third string input argument.

- The *first input argument* must be a string with the file name of the Modelica code, with Modelica file extension `.mo`. If the Modelica file is not in the current directory of Python, then the file path must also be included.
- The *second input argument* must be a string with the name of the Modelica model, including the namespace if the model is wrapped within a Modelica package.
- *A third input argument* is used if the Modelica model builds on other Modelica code, e.g. the Modelica Standard Library.

The result of using the object constructor is a Python object.

Example 1 *Use of constructor.*

Suppose we have a Modelica model with name *CSTR* wrapped in a Modelica package *Reactors* — stored in file `Reactor.mo`:

```
package Reactors
    // ...
    model CSTR
        /// ...
    end CSTR;
    //
end Reactors;
```

If this model does not use any external Modelica code and the file is located in the current Python directory, the following Python code instantiates a Python object `mod`:

[8] The Python prompt `>>>` is not typed, and does not appear in script files, in iPython or in Jupyter notebooks.

```
>>> mod = ModelicaSystem('Reactors.mo',
 'Reactors.CSTR')
```

The user is free to choose any valid Python label name for the Python object.

All methods of class *ModelicaSystem* refers to the instantiated object, in standard Python fashion. Thus, method `simulate()` is invoked with the Python command:

```
>>> mod.simulate()
```

In the subsequent overview of methods, the object name is not included. In practice, of course, it must be included in order to operate on the object in question.

Methods may have no input arguments, one, or several input arguments. Methods may or may not return results — if the methods do not return results, the results are stored within the object.

2.4.2 Utility Routines, Converting Modelica ↔ FMU

Two utility methods convert files between Modelica files with file extension `.mo` and Functional Mock-up Unit (FMU) files with file extension `.fmu`.

1. `convertMo2Fmu()` — method for converting the Modelica model of the object, say `ModelName`, into FMU file.
   - Required input arguments: none, operates on the Modelica file associated with the object.
   - Optional input arguments:
     - `className`: string with the class name that should be translated,
     - `version`: string with FMU version, "1.0" or "2.0"; the default is "1.0".
     - `fmuType`: string with FMU type, "me" (model exchange) or "cs" (co-simulation); the default is "me".
     - `fileNamePrefix`: string; the default is \'className\'.
     - `generatedFileName`: string, returns the full path of the generated FMU.
   - Result: file `ModelName.fmu` in the current directory

2. `convertFmu2Mo(s)` — method for converting an FMU file into a Modelica file.
   - Required input arguments: string `s`, where `s` is name of FMU file, including extension `.fmu`.
   - Optional input arguments: a number of optional input arguments, e.g. the possibility to change working directory for the imported FMU files.
   - Result: Assume the name of the file is `fmuName.fmu`. Then file `fmuName_me_FMU.mo` is generated in the current Python directory.

2.4.3 Getting and Setting Information

Quite a few methods are dedicated to getting and setting information about objects. With two exceptions — `getQuantities()` and `getSolutions()` — the *get* methods have identical use of input arguments and results, while all the set methods have identical use of input arguments, with results stored in the object.

Getting Quantity Information

Method `getQuantities()` does not accept input arguments, and returns a *list of dictionaries*, one dictionary for each quantity. Each dictionary has the following keys — with values being strings, too.

- `Changeable` — value `'true'` or `'false'`,
- `Description` — the string used in Modelica to describe the quantity, e.g. `'Mass in tank, kg'`,
- `Name` — the name of the quantity, e.g. `'T'`, `'der(T)'`,`'n[1]'`,`'mod1.T'`, etc.,
- `Value` — the value of the quantity, e.g. `'None'`, `'5.0'`, etc.,
- `Variability` — `'continuous'`, `'parameter'`.

When applying the *Pandas* method *DataFrame* to the returned list of dictionaries, the result is a conveniently typeset table in Jupyter notebooks. Modelica `constants` are not included in the returned quantities.

Standard Get Methods

We consider methods `getXXXs()`, where `XXXs` is in {`Continuous`, `Parameters`, `Inputs`, `Outputs`, `SimulationOptions`, `OptimizationOptions`, `LinearizationOptions`}. Thus, methods `getContinuous()`, `getParameters()`, etc.

Two calling possibilities are accepted.

- `getXXXs()`, i.e. without input argument, returns a *dictionary* with names as keys and values as ... values.
- `getXXXs(S)`, where `S` is a *sequence* of strings of names, returns a *tuple* of values for the specified names.

Getting Solutions

We consider method `getSolutions()`. Two calling possibilities are accepted.

- `getSolutions()`, i.e. without input arguments, returns a *list of strings of names* of quantities for which there is a *solution = time series*.[9]

[9] The reason why a dictionary with every name as key and time series as values is not returned, is that the amount of data would be exhaustive.

- `getSolutions(S)`, where `S` is a *sequence* of strings of names, returns a *tuple* of values = 1D numpy arrays = time series for the specified names.

Setting Methods

The information that can be set is a subset of the information that can be set. Thus, we consider methods `setXXXs()`, where `XXXs` is in {`Parameters`, `Inputs`, `SimulationOptions`, `OptimizationOptions`, `LinearizationOptions`}, thus methods `setParameters()`, `setInputs()`, etc. Two calling possibilities are accepted.

- `setXXXs(K)`, with `K` being a sequence of keyword assignments of type `quantity name = value`. Here, the quantity name could be a parameter name (i.e., not a string), an input name, etc.
  - For parameters and simulation/optimization/-linearization options, the value should be a numerical value or a string (e.g. a string of ODE solver name such as `'dassl'`, etc.).
  - For inputs, the value could be a numerical value if the input is constant in the time range of the simulation,
  - For inputs, the value could alternatively be a *list* of tuples (t_j, u_j), i.e., $[(t_1,u_1),(t_2,u_2),\ldots,(t_N,u_N)]$ where the input varies linearly between (t_j,u_j) and (t_{j+1},u_{j+1}), where $t_j \leq t_{j+1}$, and where at most two subsequent time indices t_j, t_{j+1} can have the same value. As an example, `[...,(1,10),(1,20),...]` describes a perfect jump in input value from value 10 to value 20 at time instance 1.
  - This type of sequence of input arguments does not work for certain quantity names, e.g. `'der(T)'`, `'n[1]'`, `'mod1.T'`, because Python does not allow for label names `der(T)`, `n[1]`, `mod1.T`, etc.
- `setXXXs(**D)`, with `D` being a *dictionary* with quantity names as keywords and values as described with the alternative input argument `K`.

2.4.4 Operating on Python Object: Simulation, Optimization

The following methods operate on the object, and have no *input arguments*. The methods have no return values, instead the results are stored within the object.

- `simulate()` — simulates the system with the given simulation options
- `optimize()` — optimizes the Optimica problem with the given optimization options

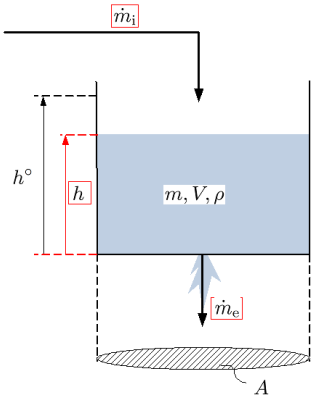


Figure 1. Driven water tank, with externally available quantities framed in red: initial mass is emptied through bottom at rate $\dot{m}_\mathrm{e}$, while at the same time water enters the tank at rate $\dot{m}_\mathrm{i}$.

To retrieve the results, method `getSolutions()` is used as described previously.

2.4.5 Operating on Python Object: Linearization

The following methods are proposed for linearization[10]:

- `linearize()` — with no input argument, returns a tuple of 2D numpy arrays (matrices) A, B, C and D.
- `getLinearInputs()` — with no input argument, returns a list of strings of names of inputs used when forming matrices B and D.
- `getLinearOutputs()` — with no input argument, returns a list of strings of names of outputs used when forming matrices C and D.
- `getLinearStates()` — with no input argument, returns a list of strings of names of states used when forming matrices A, B, C and D.

3 Use of API for Model Analysis

3.1 Case Study: Simple Tank Filled with Liquid

We consider the tank in Figure 1 filled with water.

Water with initial mass $m(0)$ is emptied by gravity through a hole in the bottom at effluent mass flow rate $\dot{m}_\mathrm{e}$, while at the same time water is filled into the tank at influent mass flow rate $\dot{m}_\mathrm{i}$.

Our *modeling objective* is to find the liquid level h. This objective is illustrated by the *functional diagram* in Figure 2.

The functional diagram depicts the causality of the *system* ("Tank with influent and effluent mass flow"), where *inputs* (green arrow) cause a change in the system and is

[10] This part of the API is not completed at the moment, and may change.

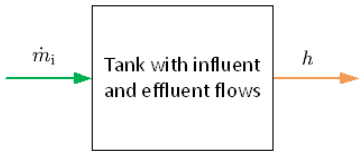


Figure 2. Functional diagram of tank with influent and effluent flow.

observed at *outputs* (orange arrow)[11]. Here, the input variable is the influent mass flow rate $\dot{m}_\mathrm{i}$, while the output variable is the quantity we are interested in, h.

3.2 Model Summary

The model can be summarized in a form suitable for implementation in Modelica as

$$\frac{dm}{dt} = \dot{m}_\mathrm{i} - \dot{m}_\mathrm{e} \tag{1}$$

$$m = \rho V \tag{2}$$

$$V = Ah \tag{3}$$

$$\dot{m}_\mathrm{e} = K\sqrt{\frac{h}{h^\nabla}}. \tag{4}$$

To complete the model description, we need to specify model parameters and operating conditions. Model parameters (constants) are given in Table 1.

The operating conditions are given in Table 2.

3.3 Modelica Encoding of Model

The Modelica code describes the core model of the tank, `ModWaterTank`, and consists of a *first section* where constants and variables are specified, and a *second section* where the model equations are specified.

```
model ModWaterTank
    // Main driven water tank model
    // author:    Bernt Lie
    //            University College of
    //            Southeast Norway
    //            April 18, 2016
    //
    // Parameters
    constant Real rho = 1 "Density";
    parameter Real A = 5 "Tank area";
    parameter Real K = 5 "Valve const";
    parameter Real h_max = 3 "Scaling";
    // Initial state parameters
    parameter Real h_0 = 1.5
    "Init.level";
    parameter Real m_0 = rho*h_0*A
    "Init.mass";
    // Declaring variables
    // -- states
    Real m(start = m_0, fixed = true)
        "Mass in tank, kg";
    // -- auxiliary variables
    Real V "Tank liquid volume, L";
    Real md_e "Effluent mass flow";
    // -- input variables
    input Real md_i "Influent mass
    flow";
    // -- output variables
    output Real h "Tank liquid level,
    dm";
// Equations constituting the model
equation
    // Differential equation
    der(m) = md_i - md_e;
    // Algebraic equations
    m = rho*V;
    V = A*h;
    md_e = K*sqrt(h/h_max);
end ModWaterTank;
```

As seen from the *first section* of model `ModWaterTank`, the model has 4 essential parameters (`rho-h_max`) of which one is a Modelica *constant* (`rho`) while other 3 are design parameters, compare this to Table 1. Furthermore, the model contains 2 "initial state" parameters, where 1 of them can be chosen at liberty, `h_0`, while the other one, `m_0`, is computed automatically from `h_0`, see Table 2. The purpose of the "free parameter" `h_0` is that it is easier for the user to specify level than mass. Also, free "initial state" parameters makes it possible for the user to change the initial states from outside of model `ModWaterTank`, e.g., from Python.

Next, one variable is given with initial value — the state `m` — is initialized with the "initial state" parameter `m_0`. Then, 2 variables are defined as auxiliary variables (algebraic variables), `V` and `md_e`.[12]

Table 1. Parameters for driven tank with constant cross sectional area.

Parameter	*Value*	*Unit*	*Comment*
ρ	1	kg/L	Density of liquid
A	5	dm^2	Constant cross sectional area
K	5	kg/s	Valve constant
h^∇	3	dm	Level scaling

Table 2. Operating condition for driven tank with constant cross sectional area.

Quantity	*Value*	*Unit*	*Comment*
$h(0)$	1.5	dm	Initial level
$m(0)$	$\rho h(0)A$	kg	Initial mass
$\dot{m}_\mathrm{i}(t)$	2	kg/s	Nominal influent mass flow rate; may be varied

[11] Although Modelica is an *acausal* modeling language, it is useful to think in terms of causality during model development.

[12] `md` is notation for m with a dot, $\dot{m}$, i.e., a mass flow rate.

One input variable is defined — `md_i` — this is the influent mass flow rate $\dot{m}_i$, see Table 2. Inputs are characterized by that their values are not specified in model the core model — here `ModWaterTank`. Instead, their values must be given in an external model/code — we will specify this input in Python. Finally, 1 output is given — `h`.

In the *second section* of model `ModWaterTank`, the Model equations exactly map the mathematical model given in Section 3.2.

For illustrative purposes, the core model `ModWaterTank` is wrapped within a package named `WaterTank` and stored in file `WaterTank.mo`,

```
package WaterTank
    // Package for simulating
    //    driven water tank
    // author:    Bernt Lie
    //            University College of
    //            Southeast Norway
    //            April 18, 2016
    //
    model ModWaterTank
        // Main driven water tank model
        // ....
        ....
    end ModWaterTank;
    // End package
end WaterTank;
```

3.4 Use of Python API

First, the following Python statements are executed — we did this in Jupyter notebook.

```
from OMPython import ModelicaSystem
import numpy as np
import numpy.random as nr
%matplotlib inline
import matplotlib.pyplot as plt
import pandas as pd
LW = 2
```

Here, we use *NumPy* to handle simulation results, etc. The random number package will be used in a sensitivity/-Monte Carlo study. The *magic* function `%matplotlib inline` is used to embed *Matplotlib* plots within the Jupyter notebook; to save these plots into files, simply right-click the plots. However, more options for saving files are available if the magic function is *excluded*, and instead command `plt.show()` is added after the plot commands have been completed. *Pandas* are used to illustrate presenting data in tables in Jupyter notebook. Finally, label `LW` is used to give a conform line width in plots.

3.5 Basic Simulation of Model

We instantiate object `tank` with the following command:

```
tank = ModelicaSystem('WaterTank.mo',
 'WaterTank.ModWaterTank')
```

```
In [9]: pd.DataFrame(q)
```

Out[9]:

	Changeable	Description	Name	Value	Variability
0	true	Mass in tank, kg	wt.m	None	continuous
1	false	Mass in tank, kg	der(wt.m)	None	continuous
2	false	External input, passed on to instantiated mode...	_md_i	None	continuous
3	false	Tank liquid level, dm	wt.h	None	continuous
4	false	Effluent mass flow rate from tank, kg/s	wt.md_e	None	continuous
5	true	Cross sectional area of tank, dm2	wt.A	5.0	parameter
6	true	Valve constant, kg/s	wt.K	5.0	parameter
7	true	Initial tank level, dm	wt.h_0	1.5	parameter
8	true	Scaling level, dm	wt.h_s	3.0	parameter
9	false	Initial tank mass, kg	wt.m_0	None	parameter
10	false	Tank liquid volume, L	wt.V	None	continuous
11	false	Influent mass flow rate to tank, kg/s	wt.md_i	None	continuous

Figure 3. Typesetting of Data Frame of quantity list in Jupyter notebook.

whereupon Python/Jupyter notebook responds that the OMC Server is up and running the file. Next, we are interested in which *quantities* are available in the model. In the sequel, Python prompt `>>>` is used when Jupyter notebook actually uses `In[*]` — where `*` is some number, while the response in Jupyter notebook is prepended with `Out[*]`.

```
>>> q = tank.getQuantities()
>>> type(q)
list
>>> len(q)
11
>>> q[0]
{'Changeable': 'true',
'Description': 'Mass in tank, kg',
'Name': 'm',
'Value': None,
'Variability': 'continuous'}
>>> pd.DataFrame(q)
```

The last command leads Jupyter notebook to typeset a tabular presentation of the quantities, Figure 3. The results in Figure 3 should be compared to the Modelica model in Section 3.3. Observe that Modelica *constants* are not included in the quantity list.

Next, we check the simulation options:

```
>>> tank.getSimulationOptions()
{'solver': 'dassl',
'startTime': 0.0,
'stepSize': 0.002,
'stopTime': 1.0,
'tolerance': 1e-06}
```

It should be observed that the *stepSize* is the frequency at which solutions are *stored*, and is not the step size of the solver. The number of data points stored, is thus `(stopTime-startTime)/stepSize` with due rounding. This means that if we increase the *stopTime* to a large number, we should also increase the *stepSize* to avoid storing a large number of information.

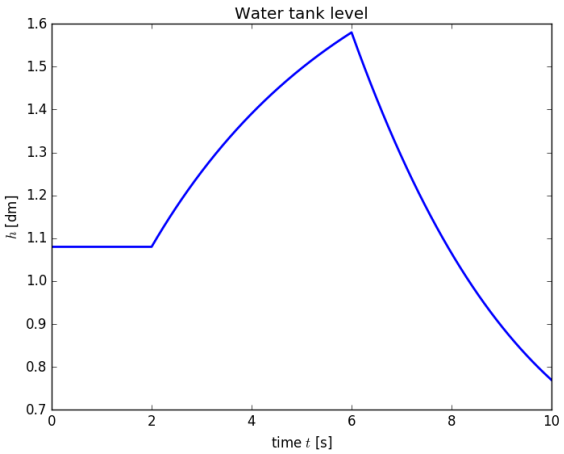


Figure 4. Tank level when starting from steady state, and $\dot{m}_{\mathrm{i}}(t)$ varies in a straight line between the points $(t_j, \dot{m}_{\mathrm{i}}(t_j))$ given by the list $[(0,3),(2,3),(2,4),(6,4),(6,2),(10,2)]$.

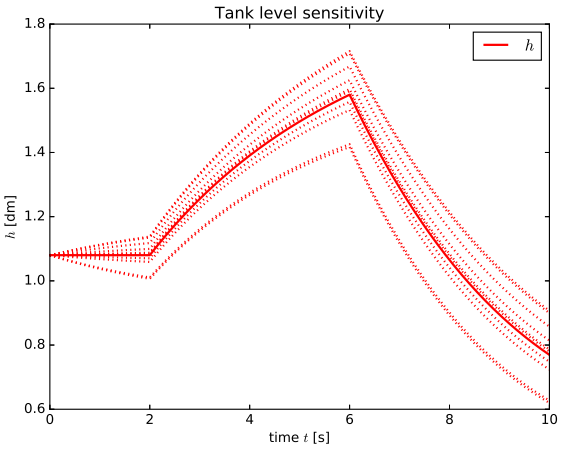


Figure 5. Uncertainty in tank level with a 5% uncertainty in valve constant K. The input is like in Figure 4.

To this end, we want to simulate the system for a long time, until the level reaches steady state. Possible inputs are:

```
>>> tank.getInputs()
{'md_i': None}
```

where value `None` implies that the available input, `md_i`, has yet not been set. We could use `None` as input, which will be interpreted as zero. But let us instead set $\dot{m}_{\mathrm{i}} = 3$, simulate for a long time, and change "initial state" parameter $h(0)$ to the steady state value of h:

```
>>> tank.setInputs(md_i=3)
>>> tank.setSimulationOptions\
    (stopTime=1e4, stepSize=10)
>>> tank.simulate()
>>> h = tank.getSolutions('h')
>>> tank.setParameters(h_0 = h[-1])
```

Next, we set back to stop time to 10, and specify an input sequence with a couple of jumps:

```
>>> tank.setSimulationOptions\
    (stopTime=10, stepSize=0.02)
>>> tank.setInputs(md_i = [(0,3),(2,3),
(2,4),(6,4),(6,2),(10,2)])
```

Finally, we simulate the model with the time varying input, and plot the result:

```
>>> tank.simulate()
>>> tm, h = tank.getSolutions('time',\
    'h')
>>> plt.plot(tm,h,linewidth=LW,
color='blue', label=r'$h$')
>>> plt.title('Water tank level')
>>> plt.xlabel(r'time $t$ [s]')
>>> plt.ylabel(r'$h$ [dm]')
```

The result is displayed in Figure 4.

3.6 Parameter Sensitivity/Monte Carlo Simulation

It is of interest to study how the model behavior varies with varying uncertain parameter values, e.g. the effluent valve constant K. This can be done as follows:

```
>>> par = tank.getParameters()
>>> K = par['K']
>>> KK = K + (nr.randn(10)-0.5)*K/20
>>> tank.simulate()
>>> tm, h = tank.getSolutions('time',\
    'h')
>>> plt.plot(tm,h,linewidth = LW,
color = 'red', label=r'$h$')
>>> for k in KK:
        tank.setParameters(K=k);
        tank.simulate()
        tm, h = tank.getSolutions\
('time','h')
        plt.plot(tm,h,linewidth=LW,
           color='red',linestyle=\
'dotted',label='_nolabel_')
>>> plt.title('Tank level sensitivity')
>>> plt.xlabel(r'time $t$ [s]')
>>> plt.ylabel(r'$h$ [dm]')
>>> plt.legend()
```

The result is as shown in Figure 5.

4 Discussion and Conclusions

This paper introduces some ongoing work on extending OpenModelica with a Python API, so that Modelica models can be run and analyzed from within Python. The new Python API is briefly described, and the use of this API is then illustrated by simulating a very simple model of a water tank.

Future work will include further testing, e.g., with optimization, extending the API so that it works on more plat-

forms, and extending the API to include analytic model linearization.

References

S. Bajracharya. *Enhanced OpenModelica Python Interface*. Master's thesis, Linköping University, Department of Computer and Information Science, 2016.

K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. Society for Industrial and Applied Mathematics, SIAM, Philadelphia, 2nd edition, 1987. doi:10.1137/1.9781611971224.

P. Fritzson. *Principles of Object-Oriented Modeling and Simulation with Modelica 3.3: A Cyber-Physical Approach, second edition*. Wiley-IEEE Press, 2014. ISBN 978-1-118-85912-4.

P. Fritzson, P. Aronsson, A. Pop, H. Lundvall, K. Nyström, L. Saldamli, D. Broman, and A. Sandholm. Openmodelica – a free open-source environment for system modeling, simulation, and teaching. In *Proceedings of the 2006 IEEE Conference on Computer Aided Control System Design*, Oct 4–6 2006.

A.K. Ganeson. *Design and Implementation of a User Friendly OpenModelica - Python interface*. Master's thesis, Linköping University, 2012.

A.K. Ganeson, F. Fritzson, O. Rogovchenko, A. Asghar, M. Sjölund, and A. Pfeiffer. An openmodelica python interface and its use in pysimulator. In *Proceedings of the 9th International Modelica Conference*, September 3-5 2012. doi:10.3384/ecp12076537.

M.A.S. Perera, T.A. Hauge, and C.F. Pfeiffer. Parameter and State Estimation of Large-Scale Complex Systems Using Python Tools. *Modeling, Identification and Control*, 36(3): 189–198, 2015a. doi:10.4173/mic.2015.3.6.

M.A.S. Perera, B. Lie, and C.F. Pfeiffer. Structural Observability Analysis of Large Scale Systems Using Modelica and Python. *Modeling, Identification and Control*, 36(1):53–65, 2015b. doi:10.4173/mic.2015.1.4.

A. Pfeiffer, M. Hellerer, S. Hartweg, M. Otter, and M. Reiner. PySimulator – A Simulation and Analysis Environment in Python with Plugin Infrastructure. In *Proceedings of the 9th International Modelica Conference*, pages 523–536, September 3-5 2012. doi:10.3384/ecp12076523. URL `http://dx.doi.org/10.3384/ecp12076523`.

Hardware-in-the-Loop Simulation for Machines based on a Multi-Rate Approach

Christian Scheifele Alexander Verl

Institute for Control Engineering of Machine Tools and Manufacturing Units, University of Stuttgart, Germany,
christian.scheifele@isw.uni-stuttgart.de

Abstract

The commissioning of the entire control system using a digital shadow of the machine offers extensive advantages in industrial control engineering for machine manufacturers and machine integrators. The growing use of a Hardware-in-the-Loop Simulation (HiLS) in the engineering process is accompanied by the steady increase in demands regarding model depth and model scope of the virtual machine. Especially in the area of material flow simulation, currently used simulation setups of HiL-Simulators reach their limits because of the limitation on a single simulation solver. This paper presents an approach on how a virtual machine could be realized based on several interlinked simulation solvers connected by a multi-rate approach to increase the model depth and model scope.

Keywords: hardware-in-the-loop simulation, co-simulation, multi-rate simulation, virtual commissioning, material flow simulation

1 Introduction

In the context of Industry 4.0, the use of digital methods and tools over the complete life cycle of a production system plays a vital role because of the increasing degree of complexity of modern production systems. In this context, one speaks of the 'virtual production': The seamless digital modelling of product installations and processes for experimentation purposes. In the area of industrial control engineering the virtual commissioning of machine tools presents great potential. Simulative methods are used more and more by machine manufacturers and machine integrators in the engineering process: In the course of virtual commissioning, the control system can be put into operation at an early development stage and before the real machine is available using a 'virtual machine'. With the aid of a virtual machine, the control system is tested regarding quality and performance. Furthermore, unforeseen errors are eliminated. In summary, virtual commissioning saves time and money and simplifies the engineering process. There is a need for further research in order to reach the vision of an encompassing virtual production.

Regarding the simulation setup of a virtual commissioning, several test configurations, such as Model-in-the-Loop (MiL), Software-in-the-Loop (SiL) and Hardware-in-the-Loop (HiL), can be distinguished. Especially in the context of CNC machines (CNC - Computerized Numerical Control), Hardware-in-the-Loop Simulation (HiLS) offers many advantages because the entire control system can be tested without any technical modifications or adaptations. In the context of CNC machines, the HiLS describes a test configuration, where the real control system is connected with a virtual machine based on a single simulation solver via the real communication periphery (Pritschow and Röck, 2004). However, in order to meet the requirements on a time-synchronous and lossless data processing in relation to the deterministic cycle time of the real control system, the machine simulation has to process the control outputs to control inputs in between the deterministic control cycle time (today 1ms for CNC machine tools). These high demands on time-deterministic algorithms cause restrictions on the model depth and model scope of a virtual machine. There are reduction schemes and numerical integration techniques available, which enable an efficient computation of the simulation models in some cases. However, computation-intensive and non-deterministic algorithms in the field of structural mechanics (e.g. real-time capable finite element models, flexible multibody systems), process simulation (e.g. chip formation), 3D-kinematic simulation with collision detection and simulation of the dynamics of material flow systems can only be used if simplified and adapted simulation models can be found. In summary, the current simulation setup based on a single simulation solver reaches its limits considering the simulation of an encompassing virtual production.

This paper presents the objective of a simulator based on several interlinked simulation solvers with different real-time requirements and cycle times connected by a multi-rate approach to increase the model depth and model scope of a HiLS of machines. The feasibility is demonstrated by the example of a material flow.

This paper is organized as follows: In section II preliminary work is presented that is considered relevant to this research. Section III describes the current

simulation setup of a HiLS in industrial control engineering and its limitations. On this basis, an expansion of the current simulation setup based on a multi-rate approach is motivated in section IV. The case study on a physics-based material flow simulation is demonstrated in section V. The paper closes with a summary and an outlook in section VI considering future work respective this research.

2 Related Work

Pritschow and Röck introduce in (Pritschow and Röck, 2004) a simulation setup of a HiLS for machine tools using a real CNC. The integration of the real CNC-System in the simulation loop requires a time-deterministic simulation of the machine that runs under a real-time operating system. An example where this simulation setup reaches its limitations regarding computation-intensive and non-deterministic simulation models is a material flow simulation using a physics-based simulation approach:

Physics-based material flow models are predestined for the simulation of the material flow dynamics of a virtual production. Zäh, Lacour and Spitzweg propose in (Zäh et al, 2008) a five step modelling process based on the CAD model that yields in a physical and kinematical model of a material flow system.

Few approaches address the integration of a physics-based material flow model into a time-deterministic virtual machine (Hoher et al, 2011; Hoher and Verl, 2012; Neher and Lechler, 2015). However, Hoher and Verl demonstrate in (Hoher and Verl, 2012) that a physics-based simulation approach is only possible with a small number of moving objects (70 dynamic objects with a simplified modeling already require up to 1,5 ms) within the described simulation setup of a HiLS.

In order to avoid these limitations, this paper presents the objective of an expansion of the simulation setup introduced in (Pritschow and Röck, 2004) by a multi-rate approach to increase the model depth and model scope. In the field of multi-rate simulation, extensive mathematical research is available (Gear and Wells, 1984) (Muttay-Smith, 1984). However, there is no solution present, which shows such a simulation setup in the field of machines and industrial control engineering. A solution needs to be found, which focus the specific scientific questions regarding the integration of a real control system in the simulation loop (HiLS), where the simulator consists of several interlinked simulation solvers.

3 HILS of Machine Tools and its Limitations

3.1 Real-Time Requirements

In a real-time simulation, the simulation time synchronizes with the real time. In the field of industrial control engineering, two different types of real-time requirements need to be distinguished with regard to used algorithms:

1. *Soft real-time requirements*: The simulation usually calculates simulation results in a timely manner. Significant deviations are rare but possible. Windows-based real-time simulation cores usually obey soft real-time requirements.
2. *Hard real-time requirements*: The simulation always calculates simulation results in a timely manner. To achieve equidistant time intervals a real-time operating system is required.

3.2 Simulation Setup for a HiLS of Machine Tools

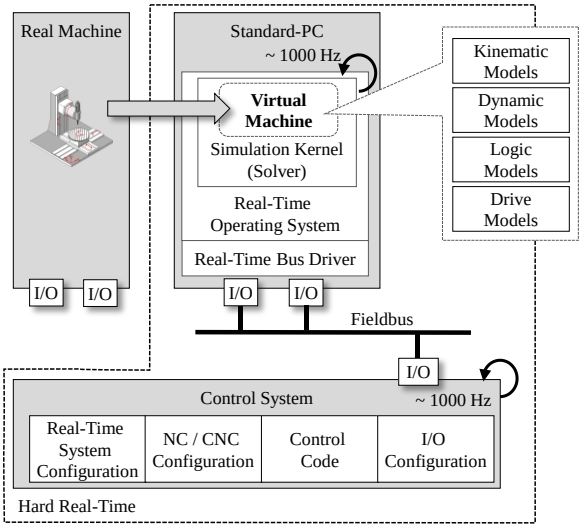


Figure 1. Simulation setup for a HiLS of machine tools using real CNC.

The decisive advantage of a HiLS of machine tools, as depicted in Figure 1, is the commissioning of an unmodified and entire control system. From the point of view of the real control system, there shell be no difference between the real and the simulated machine. To meet the requirements on a time-synchronous and lossless data processing in relation to the cycle time of the control, the simulation solver runs as real-time task on the real-time operating system (Pritschow and Röck, 2004). This ensures that time sensitive (fast changing I/Os) and time synchronous events (drive amplifier) can be handled by the simulator. Hence, the requirements on the time-deterministic simulation models are (Pritschow and Röck, 2004):

- A time-deterministic kernel of the underlying operating system to run the simulation solver
- The algorithms of the simulation models for the machine simulation must be time-deterministic
- Simulation cycle time ~ 1 ms (same as control system cycle time)

The simulated machine consists of multiple behavior models. Starting with the I/O signals on the fieldbus,

behavior models are required to simulate the single bus devices, such as a drive amplifier. Regarding the behavior model of a drive amplifier, it is important to react to every bit change inside the control word with correct status word. Downstream of the simulated bus devices, additional models such as kinematic, logic or dynamic models of the machine are needed to generate a realistic observation.

3.3 Limitations of the Simulation Setup

Following limitations can be determined:

1. *Single simulation solver*: The described simulation setup of a HiLS uses a single simulation solver running as real-time task on a real-time operating system. The setup applies an uniform cycle time (derived from the control system cycle time) and hard real-time requirements (simulation results must be guaranteed before the next control step starts) to all parts of the model. These two properties limit the model depth and model scope of the virtual machine, because of desired time-deterministic and performant algorithms and weak exploitation of available processing power.
2. *Parallelization*: Multi-core processors with an increasing number of cores as well as developments in the field of GPGPU (GPGPU - general-purpose computing on graphics processing units) are emerging since the beginning of this century offering a continuously increasing performance for simulation applications on standard PCs. A distribution of real-time tasks across different cores (Multi-core) requires a division of the simulation solver into functional units. An increase in computing performance by harnessing the power of the GPU (GPU – graphics processing unit) requires a coupling of a soft real-time simulation solver under Windows because GPGPU is currently not possible under real-time operating systems.
3. *Available simulations cores*: Simulation approaches of various simulation disciplines need to be combined to simulate the overall behavior of a machine or production. In the meantime, highly specialized real-time simulation cores for various simulation disciplines were developed. These simulation cores are mostly Windows-based and therefore impossible to run under a real-time operating system. This is a further argument for coupling Windows-based simulation cores.

The following amendments to the simulation setup are required to achieve an increasing model depth as well as an increasing model scope (see Figure 2):

- Splitting up of the single simulation solver into functional units: several interlinked simulation solvers with various cycle times
- Different real-time requirements to the simulation solvers: Enhancement of the simulation setup by soft real-time simulation solvers
- Usage of multi-rate and multi-step methods
- Integration of available highly specialized simulation cores
- Parallelization of simulation tasks: Multi-core and GPGPU support

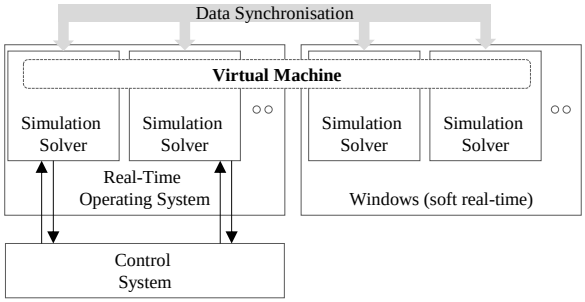


Figure 2. Multiple simulation cores.

4 Multi-Rate Simulation Approach for Machines

4.1 Multi-Rate Simulation Techniques

By splitting up of a single simulation solver into functional units with different cycle times and different real-time requirements, a synchronisation strategy is required. In the context of differential-equation models, few approaches address 'multi-rate' techniques, e.g. (Gear and Wells, 1984; Muttay-Smith, 1984).

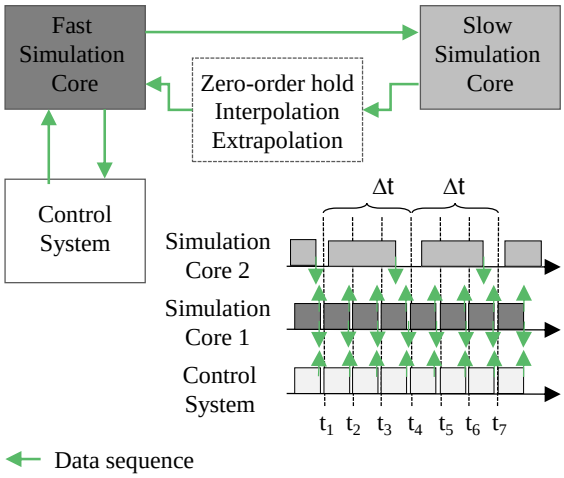


Figure 3. Multi-rate simulation.

Multi-rate methods for converting slow data sequence outputs from a slow functional unit into fast data sequence inputs for a fast functional unit can be divided into three groups, see Figure 3:

- *Zero-order hold*: Holding the last output y from the slow functional unit as input u from the fast functional unit until the output y is updated ($0 \leq a \leq 1$):

$$u_{n+a} = y_n \quad (1)$$

- *Interpolation algorithms*: Calculation of new inputs u within the range of a discrete set of known outputs y e.g. first-order interpolation ($0 \leq a \leq 1$):

$$u_{n+a} = y_n + a(y_{n+1} - y_n) \quad (2)$$

- *Extrapolation algorithms*: Estimating the value u from the values y computed beforehand e.g. first-order extrapolation ($0 \leq a \leq 1$):

$$u_{n+a} = y_n + a(y_n - y_{n-1}) \quad (3)$$

4.2 Multi-Rate Simulation Techniques for Virtual Machines

To transfer these techniques to a HiLS of machines, an application specific consideration is required:

1. *Coupling of Soft-Real-Time-Cores*: The coupling of simulation cores in soft real-time is of great interest. Soft real-time algorithms can't guarantee the calculation of simulation results if the cycle time is choosen too low. An appropriate choice of the cycle time is necessary. Therefore, the worst case need to be calculated in advance. Furthermore, a simulation step of a simulation core on Windows has to be commanded from a task on the real-time operating system.
2. *Allocation of multi-rate methods*: The choice of an available multi-rate method depends on the behavior model as well as on the characteristic of a signal. Furthermore, multiscale modeling can be considered: In a fast simulation core, a simplified and performant simulation approach is realized which is guided by a slow simulation core based on a precise simulation approach.
3. *Accuracy and stability*: Whether a behavior model requires the same cycle time as the control system is depending on the characteristic of the linked I/O signals of the control system. The effects of signal jumps as well as inaccuracies because of the multi-rate method has to be considered. Furthermore, control commands could have the demand on immediate processing by the simulation.
4. *Look-ahead simulation*: For the use of interpolation multi-rate methods, it might be necessary to parallel computing of the same simulation core with different cycle times. Furthermore, simulations faster than real-time can also be considered. As long as there is no control command, these calculations can be correct.
5. *Communication between cores*: A performant data exchange between the simulation cores is very important. Communication in-between the real-time operating system as well as between Windows and the real-time operating system has to be taken into account.

5 Case Study on a Physics-Based Material Flow Simulation

The high demands on time-deterministic algorithms is especially within the context of material flow simulation a major problem. In modern production systems, several conveyor systems combine individual machines to a material flow system. The real control system, in this case the individual machine controls as well as the superordinated production control, can be connected with a virtual production in a HiLS if a material-flow model is available. For example, looking at machines from packaging or beverage industry, the state of the control, the plant layout and operational throughput are directly related to the physical and geometric properties of the material flow. The material flow consists of a high

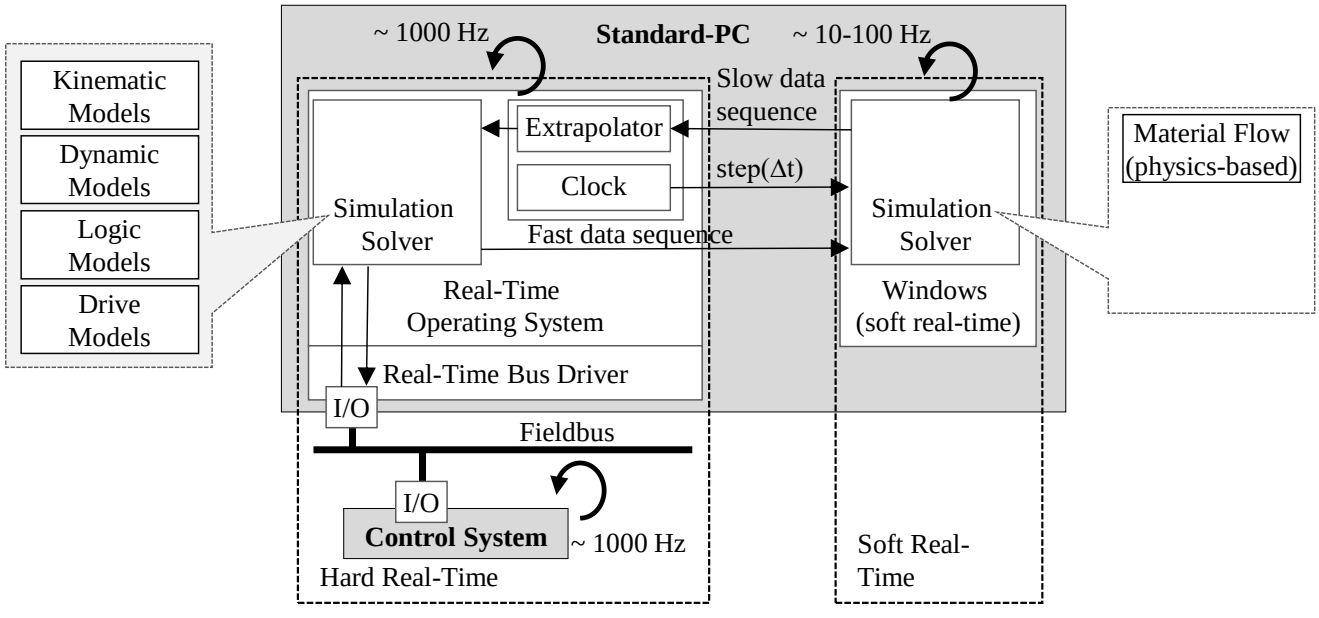


Figure 4. Simulation setup for HiLS based on a multi-rate approach of machines embedded in a material flow system.

number of moving objects (> 1000), which interact with each other over collisions. A behavior model for the dynamics of the material flow has to be found to simulate the exact process sequences and to reproduce a realistic interaction with the control system via fieldbus I/Os.

A physics-based simulation approach calculates the object motion on runtime of a simulation based on the laws of classical mechanics (rigid-body simulation). Physical (e.g. mass, coefficient of friction) and geometric (collision shapes) properties of the simulation objects as well as the object arrangement (positions and orientations) and simulation scene properties (e.g. force of gravitation) are required for modelling the initial configuration of a simulation scene. Full-grown simulation cores, called physics engines, for the simulation of physical systems are available. These physics engines provide rigid body dynamics including collision detection and can be used for a material flow simulation. However, these physics engines are not based on time-deterministic algorithms and meet only soft real-time requirements. Thus, the physics engines cannot be executed on a real-time operating system for a high number of moving objects. For coupling the physics engine, running on Windows, a suitable multi-rate method has to be found. Furthermore, the simulation core on Windows has to be commanded from a task on the real-time operating system. Figure 4 shows the simulation setup.

Regarding a conveyor belt with 245 moving cylinders (see Figure 5), the computation time of a physics-based simulation for a 40 ms time step is about 18-40 ms (see Figure 6). Thus, an appropriate method has to be found, which provides the inputs for the time-deterministic simulation solver (~ 1 ms).

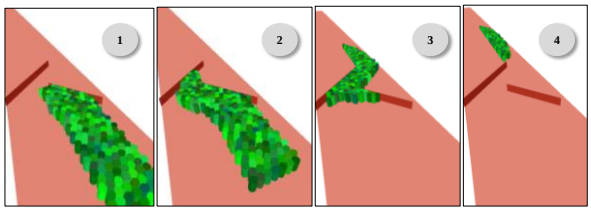


Figure 5. Material flow scene: Conveyor belt with 245 moving cylinders.

One option would be to use a zero-order hold multi-rate model. As described, this method would hold the last position vector from the slow simulation core as input for the fast simulation core until the position vector is updated.

A better option would be to use a multiscale modeling approach. A pure kinematic simulation, which lead the objects on predefined trajectories, is possible under hard real-time conditions. A kinematic simulation (in this case only translational movements) calculates in each simulation step the next position vector $\underline{r}_{n+1}$ of an object $\underline{r}_{i+1}$ by the current position $\underline{r}_n$, the current velocity vector $\underline{v}_n$ and the simulation time step Δt:

$$\underline{r}_{n+1} = \underline{r}_n + \underline{v}_n \cdot \Delta t \qquad \underline{r}_{i+1} = \underline{r}_i + \underline{v}_i \cdot \Delta t \qquad (4)$$

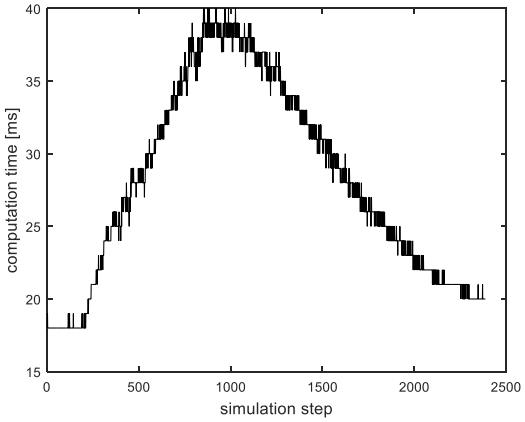


Figure 6. Computation time of the physics-based simulation.

On this basis, a multi-rate method can be developed: At n = 0, the physics-based simulation model, which is running on the slow simulation core, sends the current precise position vector $\underline{r}_p$ and velocity vector $\underline{v}_p$ of the object. The multi-rate method now updates the current position vector (2) and uses the velocity vector to generate new position vectors (3) until a new data set arrives from the slow simulation core:

$$\underline{r}_{n+0} = \underline{r}_p \qquad (5)$$

$$\underline{r}_{n+a} = \underline{r}_n + \underline{v}_n \cdot \Delta t \qquad (6)$$

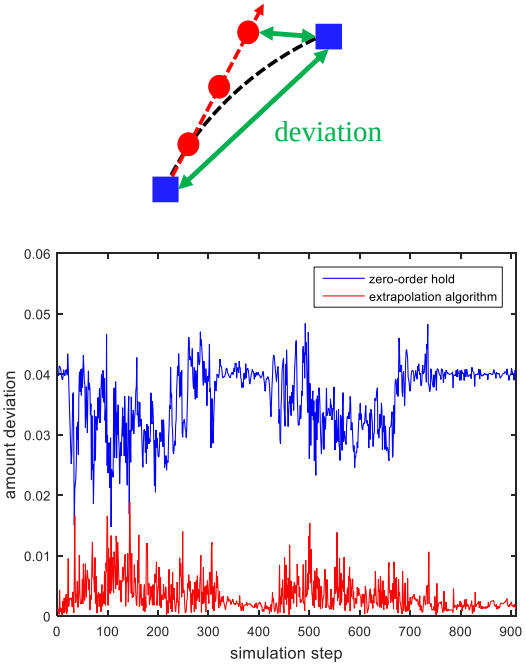


Figure 7. Amount deviation with extrapolation algorithm.

In this example, the soft real-time simulation solver with the physics-based simulation runs with a simulation cycle time of 40 ms and the hard real-time simulation solver runs with a time-deterministic cycle time of 1 ms. Compared with a zero-order hold multi-

scale method, the usage of the multiscale modeling approach enhances the model accuracy, see Figure 7. It can be seen that the accuracy of a coupled physics-based simulation is sufficient for the requirements of a HiLS of machines.

The case study is implemented in a real-time environment using Beckhoff TwinCAT real-time extension for Microsoft Windows 7 64bit. For the physics-based simulation, the physics engine NVIDIA PhysX SDK 3.3.3 and for the virtual machine the simulation tool ISG-virtuos is used.

6 Summary and Outlook

The requirements regarding model depth and model scope of the virtual machine in a HiLS are continually on the rise. The currently used simulation setup reaches its limits because of the limitation on a single simulation solver. To master these demands the division of the single simulation solver into functional units based on multi-rate methods should be considered, so that multi-core and GPGPU technologies as well as a wider range of simulation algorithms with different real-time requirements and cycle times can be used to achieve the given objective of a virtual production. Therefore a precise observation is necessary, to keep meeting the requirements on a time-synchronous and lossless data processing in relation to the cycle time of the control system.

Out of this context, this paper presents the objective of a simulator based on several interlinked simulation solvers with different real-time requirements and cycle times connected by a multi-rate approach to increase the model depth and model scope of a HiLS of machines. The feasibility is demonstrated by the example of a material flow simulation.

The objective and future work is an in depth analysis of the described approach and the analysis of the interaction with an industrial control system.

References

C. W. Gear and D. R. Wells: “Multirate linear multistep methods,” *BIT Numerical Mathematics*, 24(4):484.502, 1984. doi: 10.1007/BF01934907

S. Hoher, P. Schindler, S. Göttlich, V. Schlepper and S. Röck. System Dynamic Models and Real-time Simulation of Complex Material Flow Systems. *4th International Conference on Changeable, Agile, Reconfigurable and Virtual Production (CARV11)*, 316-321, Springer, 2011. doi: 10.1007/978-3-642-23860-4_52

S. Hoher and A. Verl. A Multi Simulator Approach of Material Flow Systems for Virtual Commissioning. *SPSIPCDRIVES 2012 Tagungsband*, G. Frey, W. Schumacher, A. Verl, 387-396, VDE, 2012.

D. J. Murray-Smith. Modelling and simulation of integrated systems in engineering: issues of methodology, quality, testing and application. *Elsevier*, 2012.

P. Neher and A. Lechler. Using game physics engines for hardware-in-the-loop material flow simulations: benefits, requirements and experiences. *Advanced Intelligent Mechatronics (AIM)*, 2015 IEEE International Conference on. IEEE, 2015. doi: 10.1109/AIM.2015.7222670

G. Pritschow and S. Röck. Hardware in the Loop Simulation of Machine Tools. *CIRP Annals-Manufacturing Technology*, 53(1):295-298, 2004. doi:10.1016/S0007-8506(07)60701-X

M. F. Zäh, F. Lacour and M. Spitzweg. Application of a physical model for the simulation of the material flow of a manufacturing plant. *IT – Information Technology*, 50(3):192-198, 2008.

Powertrain Model Assessment for Different Driving Tasks through Requirement Verification

Anders Andersson[1] Lena Buffoni[2]

[1]Swedish National Road and Transport Research Institute, Sweden, anders.andersson@vti.se
[2]IDA, Linköping University, Sweden, lena.buffoni@liu.se

Abstract

For assessing whether a system model is a good candidate for a particular simulation scenario or choosing the best system model between multiple design alternatives it is important to be able to evaluate the suitability of the system model. In this paper we present a methodology based on finite state machine requirements verifying system behaviour in a Modelica environment where the intended system model usage is within a moving base driving simulator. A use case illustrate the methodology with a Modelica powertrain system model using replaceable components and measured data from a Golf V. The achieved results show the importance of context of requirements and how users are assisted in finding system model issues.

Keywords: system model assessment, requirement modelling, Modelica, finite state machine, powertrain validations

1 Introduction

With the increasing complexity of cyber-physical systems, determining whether a particular system design alternative fulfils all the requirements that are imposed on the system under development can no longer be done manually and requires formalizing the requirements into some computable form. Verifying the validity of a system design through simulation will reduce the risk of modelling errors and allow to evaluate the suitability of the model for a particular purpose.

In the context of this paper, we illustrate the validation process on a powertrain model with an intended use in a driving simulator. A simulator model is validated before conducting a driving simulator study as well as over the whole evaluation time period and in particular whenever a developer changes parts of a model, to guarantee that the model is suitable for the intended driving tasks.

One common way to test a powertrain is to use driving cycles. For the use case in this paper we have logged a driver for two different driving cycles, the Artemis Road Driving Cycle and the 130 km/h variant of the Artemis Motorway Driving Cycle. Using this logged data it is possible to run the model offline and we use these datasets to verify that the model is working as intended using requirements.

The physical model and the requirement model are both written in Modelica (Modelica Association, 2014). Using the same language to express both the requirement and the design model simplifies the co-simulation of the two. The declarative nature of Modelica lends itself well to the description of the requirement model and the component based nature of the verification framework allows to quickly create different configurations for testing.

The paper is organized as follows: Section 2 presents the use case that will be used to illustrate the methodology, Section 3 describes the requirement model, Section 4 illustrates the setup for the whole verification framework, used in Section 5 to show the model validation process, and finally the conclusion and future work are discussed in Section 6.

2 Use Case

To acquire data for the powertrain system models and the requirement model the vehicle propulsion laboratory at Linköping University was used. In this laboratory chassis dynamometers are connected to the wheel hubs of the test vehicle, in this case a Golf V, measuring signals such as the torque and rotational speed at the driving wheels. Used setup, shown in Figure 1, is further described in (Öberg *et al*, 2013).

Figure 1. The Golf V used for measuring powertrain data connected to chassis dynamometers at the vehicle propulsion laboratory at Linköping University.

2.1 Driving Task Specification

The system models are used to simulate the powertrain during various driving tasks. Examples of such driving tasks are driving monotonously on a motorway with low traffic or city driving which typically includes more accelerations and driver input. Driving tasks can be represented by driving cycles. Thus, to connect our conclusions to real driving, two different driving cycles were used, the 130 km/h variant of the Artemis Motorway Driving Cycle and the Artemis Road Driving Cycle, see (Andre, 2004). In the laboratory, the driver was asked to drive according to the chosen driving cycle as close as possible.

Gathered data was used to parameterize the system model and also as a test case to evaluate the requirement model's performance. Since the requirement model should ensure that the system model captures the driving cycle characteristics and thus it is suitable for the represented driving task.

3 Requirement Model

Alongside the system model for the powertrain, we define the requirement model. It is important to note that the requirement model should not impact the physical model of the system and therefore has read-only access to the information necessary for the verification.

3.1 Requirement Modelling in Modelica

To represent requirements in Modelica, we use the following conventions (Schamai *et al*, 2014):

- A requirement is identified by extending the partial Requirement interface.
- A requirement is associated with a status and a set of properties to reason on the status.

A status can take the following values:

- `VIOLATED` when the conditions of the requirement are not fulfilled by the design model;
- `NOT_VIOLATED` when the conditions of the requirement are fulfilled by the design model;
- `NOT_APPLICABLE` when the requirement does not apply, for instance a requirement that describes the behavior of a vehicle when switching gears cannot be verified in a scenario where the vehicle is always in first gear. This is important to identify requirements that were never tested during a simulation.

It is important to note that the status of a requirement evolves over time, and that the status of the requirement at the last instant of the simulation cannot be used to determine whether the requirement has been violated earlier in the simulation. For this reason, each requirement is also associated with the following variables:

- `hasBeenVerified` indicates if a requirement has ever been checked during a simulation run
- `hasBeenViolated` indicates if a requirement has been violated during a simulation run

These predicates can be used to analyze the simulation results.

There is no unique way to specify a requirement model, but in this paper we choose to represent requirement as finite state machines (Thiele *et al*, 2015) because this allows to intuitively map the state of the system through inputs to the 3 possible states of the requirement. Other alternatives would have been representing requirements as conditional equations or using a dedicated library (for example (Otter *et al*, 2015)).

In is important to note that when modeling requirements as state-machines the clock frequency is a key design choice. In the current implementation the transitions are triggered by a clock event and thus a zero crossing is not detected. As a consequence, an event with a frequency shorter than the clock interval can possibly be missed. For the simulations in this study case we set the clock period to 0.5.

3.2 Powertrain Use Case Requirements

For this case study we have selected a set of 4 different requirements where three of them are related to system model validity and one is related to model logics. A textual description of the requirement set is given in Table 1.

Table 1. Textual requirement description.

Requirement ID	Description
`req1`	The accelerator and clutch pedal value should always be between 0 and 1.
`req2`	When driving calmly there should be a limited amount of gear changes per minute when using an automatic gearbox.
`req3`	When the car is moving the difference between modeled and measured engine RPM should be below an acceptable error when the clutch is not used.
`req4`	When the car is moving the difference between modeled and measured vehicle speed should be small over time.

The textual descriptions of requirements in Table 1 can be ambiguous. For instance it is unclear, whether "between 0 and 1" is an open or a closed interval. Formalizing these requirements as a computable model removes such ambiguities. As described in the previous

section, each requirement in Table 1 is modeled by a finite state machine. For an example of how these requirement finite state machines look, see Figure 2.

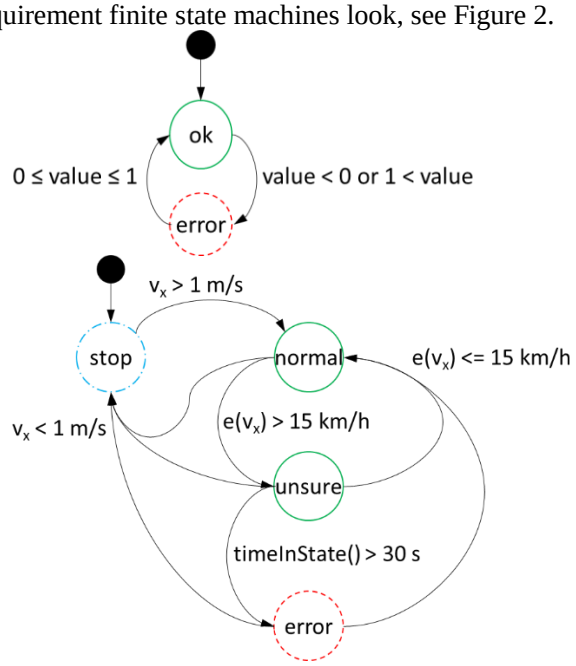


Figure 2. Finite state machines used to model **`req1`** above and **`req4`** below. The requirement status corresponds to state circle format where a green line means **`NOT_VIOLATED`**, a red dashed line means **`VIOLATED`** and a blue dash dotted line means **`NOT_APPLICABLE`**.

One of the advantages of using a component-based language to model requirements is the possibility of hierarchically composing smaller requirements into more complex requirements. For instance, `req1` is a combination of two requirements: one requirement for the acceleration pedal and another requirement for the clutch pedal. These are requirements on the inputs of the system and are grouped together for convenience purposes.

The model below is the description of the state machine for `req4` in Modelica. It corresponds to the second state machine in Figure 2. We can see here that the requirement inherits from the partial model Requirement and the different states of the requirement are represented by the states of the state machine, e.g. when the vehicle is stopped, the requirement cannot be verified and is therefore `NOT_APPLICABLE`.

```
model VehicleSpeed
  extends Requirement;
  Modelica.Blocks.Interfaces.RealInput vx;
  Modelica.Blocks.Interfaces.RealInput vx_ref;
  inner Integer y;
  Real e_vx;

  block VehicleStopped
    outer Modelica.Blocks.Interfaces.IntegerOutput y;
  equation
    y = ReqStatus.NOT_APPLICABLE;
  end VehicleStopped;
  VehicleStopped stop;

  block NormalDriving
    outer Modelica.Blocks.Interfaces.IntegerOutput y;
  equation
    y = ReqStatus.NOT_VIOLATED;
  end NormalDriving;
  NormalDriving normal;

  block DetectedError
    outer Modelica.Blocks.Interfaces.IntegerOutput y;
  equation
    y = ReqStatus.NOT_VIOLATED;
  end DetectedError;
  DetectedError unsure;

  block PersistingError
    outer Modelica.Blocks.Interfaces.IntegerOutput y;
  equation
    y = ReqStatus.VIOLATED;
  end PersistingError;
  PersistingError error;

equation
  status = y;
  e_vx = abs(vx_ref - vx);
  initialState(stop);
  transition(stop,normal,vx >= 1);
  transition(normal,stop,vx < 1,
    immediate=false,reset=true,
    synchronize=false,priority=1);
  transition(unsure,stop,vx < 1,immediate=false);
  transition(error,stop,vx < 1,immediate=false);
  transition(normal,unsure,e_vx >= 15/3.6,
    priority=2,immediate=false);
  transition(unsure,normal,e_vx <15/3.6,
    priority=2,immediate=false);
  transition(unsure,error,timeInState() >= 30,
    priority=3,immediate=false,reset=true,
    synchronize=false);
  transition(error,normal,e_vx <15/3.6,
    priority=2,immediate=false);
end VehicleSpeed;
```

The requirement model is encapsulated in a single Modelica component (the green box in Figure 3), which is then connected with the rest of the verification setup.

As requirements are parameterizable, the same requirement can be instantiated several times in a requirement model. For instance, the requirement `WithinLimits` used to verify that an input stays within certain boundaries is used twice in `req1` both for the clutch and the accelerator pedal values.

4 Verification Model

A verification model (Schamai *et al*, 2014) contains:

- The design alternative chosen to model the system
- A requirement model
- A particular scenario used for the verification

For this case study we test the set of requirements presented in Table 1 on two different versions of a powertrain model to test how each model performs and whether it is suitable for the target application. The general setup framework is shown in Figure 3.

With this setup we can thus verify different configurations of the model against different scenarios represented by different logged data. The requirement model is in this setup not changed. In Figure 3 we can also see the parts of the physical system representing the vehicle. These models will not be changed and thus in our test the same test vehicle is used, e.g. vehicle mass is not changed between different simulation setups.

In this paper we test 2 possible design alternatives in 2 scenarios for the set of requirements defined in Section 3, resulting in the total of 4 verification models.

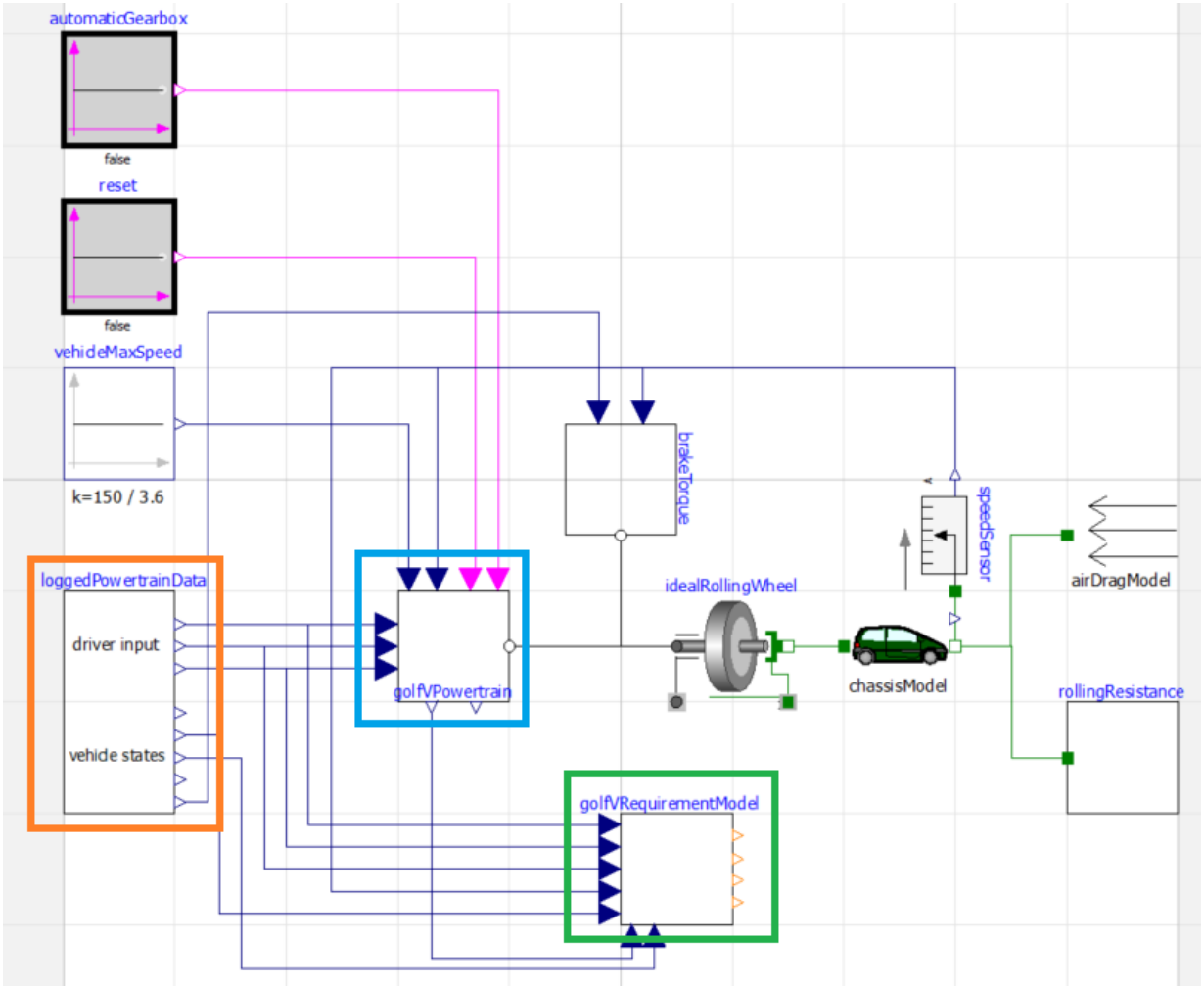


Figure 3. The complete model in OpenModelica with replaceable parts. The orange box marks the driver input which is modified depending on which data needs to be tested. The blue box is the model under test and is this is where the two different powertrain models are tested. The green box is the requirement model.

4.1 Design Alternatives

When testing different design alternatives, it is important to be able to interchange different parts of the system i.e. powertrain model easily. In Modelica a structured way to do this is to use replaceable components (Modelica Association, 2014), (Fritzson, 2014). Parts of the verification model are declared as replaceable and are instantiated with different types of components when the verification scenarios are generated.

To verify our approach we have chosen to compare two different powertrain models. These two models were created by two students and for information on the powertrain model version 1 (v1) see (Andersson *et al*, 2016) and on the powertrain model version 2 (v2) see (Zetterlund, 2015). The second model is meant to be an improved version of the first model. One of the major improvements is better performance for lower gear driving.

One of the goals in this case-study is to check whether model v2 is a more accurate model.

4.2 Verification Scenarios

To verify the powertrain system models two driving cycles were used. These are the 130 km/h variant of the Artemis Motorway Driving Cycle and the Artemis Road Driving Cycle. Data from these driving cycles are collected from the chassis dynamometers and the vehicle CAN bus in the vehicle propulsion laboratory.

Using logged data from the chassis dynamometers the powertrain models should produce the same or similar vehicle response as in the vehicle in the chassis dynamometers. Since the used driving cycle then highly influence the test of the vehicle we have chosen to use two easier tests which will test the powertrain under calm conditions.

The Artemis Motorway Driving Cycle is calm driving with one part in the middle of the test with a reduction in speed. The Artemis Road Driving Cycle is a more dynamic test and excites more dynamics in the powertrain model where for example there are three

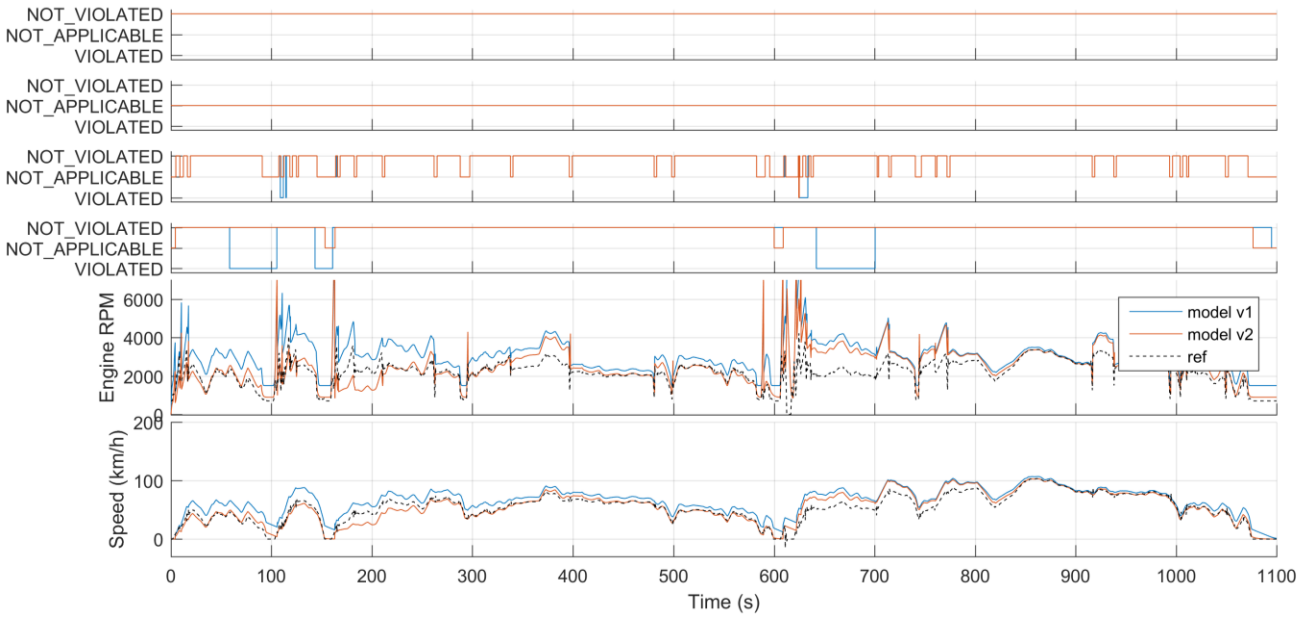


Figure 4. The powertrain model v1 and v2 during the 130 km/h variant of the Artemis Motorway Driving Cycle. The top four figures are the requirements one to four and the lower to figures show the difference in engine RPM end vehicle speed.

stops to zero vehicle speed. This means that if a powertrain model is accurate enough for a driving task when we verify our model the conclusion should be that if the requirements are passed for the Artemis Motorway Driving Cycle the model can be used for simulator studies when motorway driving is tested.

In our scenario the logged data is part of the simulation which means that a requirement can be checked during simulation and thus cancel the simulation if a requirement is violated.

Our goal is to assess whether the models can be used with the represented scenarios.

5 Simulation and Results

When running the different setups a Python interface to OpenModelica (Ganeson *et al*, 2012) is used and a Python script is used to instantiate every chosen combination of the different powertrain system models with the driver input. Each combination is associated with a unique identifier and a .mat file containing the simulation results for each combination. The data is then analyzed in Matlab to check if the requirements are violated and/or verified.

When running the simulation with a clock period of 1.0 s, we miss some events that lead to the violation of `req3`. Therefore, we use a clock period of 0.5 s.

We start by looking at the 130 km/h variant of the Artemis Motorway Driving Cycle which represents motorway driving. In Figure 4 version 1 and 2 of the powertrain system model with requirements are plotted (`req1` at the top with the other requirements consecutively downwards). From the figure we can see that the requirement on the inputs (`req1`) is never violated. This is predictable since we have been in control of the measurement equipment and has thus made sure that correct inputs have been used. The second requirement is not verified which is also expected since a manual gearbox was used during each simulation and this requirement would only be applicable for an automatic gearbox. Looking at requirements three and four (`req3` and `req4`) we see that during the time 0 to 100 when the model error is the largest for both vehicle speed and engine RPM is also when the requirements are violated for the model v1. This is illustrated more in detail in Figure 5 where the first 55 seconds is shown.

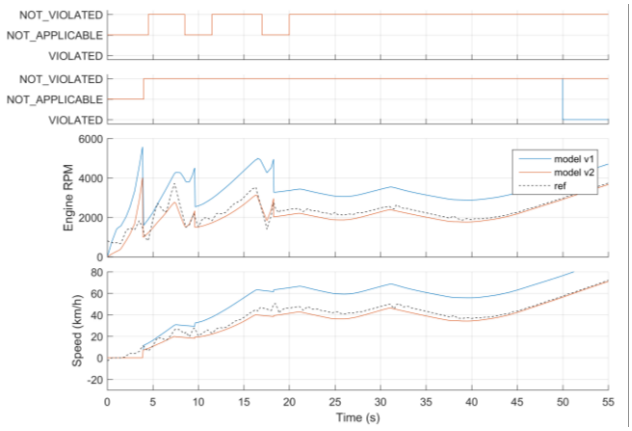


Figure 5. A close up of the first 55 seconds for models v1 and v2 during the 130 km/h variant of the Artemis Motorway Driving Cycle. Here **`req3`** and **`req4`** are shown.

In Figure 5 it can be seen that every time the clutch is pressed (at approximately the times 4, 9 and 17 seconds) an error in vehicle speed is introduced. For the system model v1 the error persists for more than 30 seconds after the third usage of the clutch and thus `req4` is

violated, see finite state machine in Figure 2. For the system model v2 it can be seen that a vehicle speed error is still present but significantly smaller. As the clutch error is below an acceptable limit `req4` is not violated. This is where model v2 is improved and thus the results are expected.

Running the simulation for our second driving cycle in the same way we get a matrix of requirements that we check for the values of `hasBeenVerified` and `hasBeenViolated`. The results are summed up in Table 2 where a green box means that the requirement has been verified and is not violated, a red box indicates that the requirement has been violated and a blue box means that the requirement has not been verified.

Table 2.

Table 2. Requirement status for the two powertrain models with two driving cycles.

		req1	req2	req3	req4
Model v1	*ArtemisMw130*				
	ArtemisRoad				
Model v2	*ArtemisMw130*				
	ArtemisRoad				

As can be seen in Table 2 the requirement set for the improved version of the model is not violated for the calmer motorway driving cycle while both models need improvement to meet the requirements for the rural road conditions. We can also see that `req1`, `req3` and `req4` have all been verified for all test cases and since automatic gearbox were never used `req2` were never verified. This result is predictable since model v2 is an improved version and should thus have better performance than model v1. From Figure 5 we can see that the difference between the models is in the region where lower gears are used. This was a region where model v1 was improved which further gives confidence that the results are correct. We also see that based on our requirement model none of the models are fit for rural road driving. This may mean that either the requirements are too strict or the model needs improvement.

Another aspect is that we do not distinguish between requirements violated early in the cycle and later on. Further inspection of the data shows that the requirements are typically violated when accelerating the vehicle from stand still to approximately 50 km/h. However, for a rural drive where the speed is above this limit most of the time this model could be used. This underlines the importance of context when looking at model assessment for a particular application. Thus a further repository with driving cycles would improve the accuracy of model verification.

Another advantage of this verification framework is that we could have several system models which doesn't violate any of the requirements for a driving task. In such a case where several models are good enough a model can be chosen based on a particular criteria not related to model accuracy. One example of such an application is models for real-time simulation where a developer can try out different numerical solvers for different system models to find the system model with good enough performance and lowest computation effort.

It should also be noted that we know that the clutch dynamics are not properly modeled, since the logged data from the CAN bus are 1 or 0. This has been taken into account in the requirements where e.g. the engine RPM is not checked when the clutch is pressed. This is also something that needs to be improved and if this is taken into account both models will violate req3. To improve the clutch model it would be good to run the vehicle while also logging intermediate clutch values to get the missing data.

6 Conclusions and Future Work

In this paper we have

- illustrated the use of requirement modeling using finite state machines in Modelica
- presented a setup for model validation of a physical system model using replaceable components to easily build different design and validation alternatives
- discussed the simulation results for different combinations of validation scenarios and used them to reason on the fitness of the chosen design alternatives for a particular purpose

In this paper we present the results at requirement level to see whether each requirement is fulfilled by a particular system design. For violated requirements however, it is not so obvious to trace the root of the problem back to the time when the requirement is violated, as the violations at time t can be due for instance to driver actions at time (t-n)., e.g. see `req4` where an error needs to persist for over 30 seconds to be `VIOLATED`. Therefore, once a violation is detected, a detailed analysis of the simulation results in the time frame surrounding the violation is still necessary by a human expert. However the semi-automation of the requirement verification process helps pinpoint the places where human intervention is necessary.

In the case-study presented in this paper, the overhead added by the simulation of the requirement model alongside the system model is considered to be negligible, however in larger-scale cases this will need to be taken into consideration. Therefore the next step is to take the overhead in account in the verification process.

Acknowledgments

This work is partially supported by the EU INTO-CPS project and the ITEA MODRIO and OPENCPS projects via the Swedish Government (Vinnova) and the German and French Government.

References

Anders Andersson, Sogol Kharrazi, Simon Lind, and Andreas Myklebust. Parameterization procedure of a powertrain model for a driving simulator. *Advances in Transportation Studies*, 2016, 1.

Michel Andre, The ARTEMIS European driving cycles for measuring car pollutant emissions. *Science of the Total Environment* 334– 335, pages 73–84, 2004.

Modelica Association. *Modelica 3.3 revision 1 specification*. 2014. URL www.modelica.org.

Peter Fritzson. *Principles of Object Oriented Modeling and Simulation with Modelica 3.3: A Cyber-Physical Approach*. 1250 pages. ISBN 9781-118-859124, Wiley IEEE Press, 2014.

Anand Ganeson, Peter Fritzson, Olena Rogovchenko, Adeel Asghar, Martin Sjölund, and Andreas Pfeiffer. An OpenModelica Python interface and its use in pysimulator. In Martin Otter and Dirk Zimmer, editors, *Proceedings of the 9th International Modelica Conference*. Linköping University Electronic Press, September 2012.

Per Öberg, Peter Nyberg, and Lars Nielsen. A new chassis dynamometer laboratory for vehicle research. *SAE International Journal of Passenger Cars- Electronic and Electrical Systems*, 2013, 6(1):152–161.

Martin Otter, Nguyen Thuy, Daniel Bouskela, Lena Buffoni, Hilding Elmqvist, Peter Fritzson, Alfredo Garro, Audrey Jardin, Hans Olsson, Maxime Payelleville, Wladimir Schamai, Eric Thomas, and Andrea Tundis. Formal requirements modeling for simulation-based verification. In Peter Fritzson and Hilding Elmqvist, editors, *Proceedings of the 11th International Modelica Conference*. Modelica Association and Linköping University Electronic Press, September 2015.

Wladimir Schamai, Lena Buffoni, and Peter Fritzson, An Approach to Automated Model Composition Illustrated in the Context of Design Verification. *Journal of Modeling, Identification and Control*, Volume 35- 2, pages 79—91, 2014.

Bernhard Thiele, Adrian Pop, and Peter Fritzson. Flattening of modelica state machines: A practical symbolic representation. In Peter Fritzson and Hilding Elmqvist, editors, *Proceedings of the 11th International Modelica Conference*. Modelica Association and Linköping University Electronic Press, September 2015.

Olof Zetterlund. Optimization of Vehicle Powertrain Model Complexity for Different Driving Tasks. *Master's thesis*, Linköping University, LiTH-ISY-EX–15/4897–SE, 2015.

Analytical Approximations and Simulation Tools for Water Cooling of Hot Rolled Steel Strip

Aarne Pohjonen [1] Vesa Kyllönen [2] Joni Paananen [1]

[1]Materials and Production Technology, University of Oulu, Finland,
Aarne.Pohjonen@Oulu.fi, Joni.Paananen@oulu.fi
[2]Technical Research Centre of Finland, VTT, Finland, Vesa.Kyllonen@vtt.fi

Abstract

Analytical approximations that can be used together with the numerical codes to obtain estimates on the temperature distribution inside of the cooled steel strip/plate are discussed. While numerical simulations can give accurate answer after the time required for calculations, the analytical approximations show how thickness and cooling rate affect the temperature distribution. We also present development of graphical user interface interaction with numerical codes for the use in designing and tuning of water cooling schedule for hot rolling strip and plate mill. Interaction of the numerical codes with user friendly frontends have been developed for the following tools: a heat conduction simulation tool for hot strip mill, a tool for calculating phase transformations for user defined cooling paths and a tool for calculating the required cooling water to cool a steel strip to a desired temperature. The functionality and interaction of the tools with the numerical codes is described.

Keywords: analytical approximation, simulation, graphical user interfaces, steel, steel processing, heat Treatment, cooling

1 Introduction

Simplified analytical approximations provide an useful way for quickly estimating physical processes and interpreting results of detailed simulations or complicated theory. The complexity of conduction of heat inside of a steel strip, taking in to account the heat released in phase transformations, the effect of transformations and temperature on the heat conductivity, heat capacity and density make the exact solution viable only through numerical calculations. However, simplified analytical results can be obtained when suitable approximations are made. With the aid of equations presented in this article, it is easy to understand how the cooling rate on the surfaces of the steel strip/plate and the thickness of the strip/plate affect the temperature distribution inside of the strip/plate.

Graphical user interface (GUI) provides other users a possibility of using advanced numerical codes without the need to spend time learning the specific, often cumbersome textual syntax, of a given code. GUI makes it easier to share the knowledge of the scientists or engineers to the user, who does not need to know the exact inner workings of the complex numerical code or analytical theory in order to benefit from the calculations in his/her work. The real industrial data can be automatically read in to the code and the user can easily experiment with the data by changing parameters from the interface. In this way the user can quickly compare numerical simulation data and experimental data to his/her experience and hypotheses. Such interaction shortens the development time from idea to experimentation as well as the number of experiments required, as many things can be experimented easily with numerical codes. Also the development of the code and theory becomes easier when the results can be compared easily to actual measured values obtained either from an industrial process or from experiments.

As part of the current development of simulation tools for the microstructural development of hot rolled steels, we have constructed graphical interfaces for three different tools. A numerical heat conduction tool (Pyykkönen et al. 2010; Pyykkönen et al. 2012a; Pyykkönen et al. 2012b; Pyykkönen et al. 2013) has been connected to a web based graphical user interface that receives real industrial data and allows the user to change the input from the data. A phase transformation model, which is currently under further development, is connected to GUI that allows user to define a cooling path by clicking with a mouse on a time-temperature diagram and calculate the phases transformed when the steel is cooled along this path. Also, GUI is created for a tool that allows user to define desired temperature path for cooling of steel and calculates the required water usage to cool the steel to this temperature.

through SIMP Show Case 2.2 program.

2 Analytical approximations for the temperature distribution during cooling

To find an estimate for the temperature distribution for a given cooling rate, we seek a time-asymptotic solution T_A to the heat equation with linear cooling rate specified at the top ($x=L$) and the bottom ($x=-L$) of a strip/plate, and discuss the time required for the material to converge to this temperature distribution, and the transient temperature distribution from initial distribution to the time-asymptotic distribution.

The time dependent temperature distribution within material can be solved from the heat equation

$$\rho c \frac{\partial}{\partial t} T(\vec{x},t) - \nabla \cdot (\kappa \nabla T(\vec{x},t)) = \sigma, \qquad (1)$$

where ρ is density, c is the specific heat capacity, κ is the heat conductivity, $\sigma(x,t)$ is amount of heat generated per time unit and $T(\vec{x},t)$ is the temperature distribution. The heat conduction within a steel strip can be approximated as 1-dimensional heat conduction, except near the edges of the strip, with heat transfer or temperature described at the top and bottom boundaries of the strip.

In order to obtain simplified analytical expression for the temperature distribution, we additionally assume that the strip is homogenous, the heat conduction, density and heat capacity are constants, and the heat release from the transformations can be neglected. This condition applies for the cooling before and after the transformations, and also when the transformation rate or heat released from given transformation is low. Heat release could be in principle estimated by adding a term corresponding to the heat release to the solution described here, but this is not considered in this work. With these approximations, the 1-dimensional heat equation is described by

$$\frac{\partial}{\partial t} T(x,t) - k \frac{\partial^2}{\partial x^2} T(x,t) = 0, \qquad (2)$$

where $k = \kappa / (\rho c)$.

Assume that $T_0(x)$ is the initial temperature distribution inside of the strip/plate. The full solution $T(x,t)$ to (2) can be written as a sum of the asymptotic solution $T_A(x,t)$ and a transition function $T_T(x,t)$ that describes the transition from initial temperature distribution to the time-asymptotic solution, as described by

$$T(x,t) = T_A(x,t) + T_T(x,t). \qquad (3)$$

We consider the case that the constant cooling rate is specified on the top and at the bottom of the strip/plate with boundary conditions described by (for generality),

$$T(-L,t) = c_0 + c_1 t, \qquad (4)$$

$$T(L,t) = d_0 + d_1 t, \qquad (5)$$

Where c_1 and d_1 are negative for cooling and positive for heating. It can be verified by direct substitution that the time-asymptotic Taylor series solution, T_A to (2) with boundary conditions given by (4) and (5) is described by

$$T_A(x,t) = \frac{(c_1+d_1)}{2} t + \frac{c_0+d_0}{2} - \frac{(c_1+d_1)L^2}{4k} + \left[\frac{d_1-c_1}{2L} t + \frac{d_0-c_0}{2L} + \frac{(c_1-d_1)L}{12k} \right] x + \frac{c_1+d_1}{4k} x^2 + \frac{d_1-c_1}{12kL} x^3. \qquad (6)$$

This solution corresponds to case where the material has been cooled long enough, with constant rates on top and bottom of the strip/plate, in order for the distribution to relax to it (there is only linear time dependence in the solution and boundary conditions).

The solution can be further simplified if top and bottom cooling rates are the same and both ends are at the same temperature when t = 0, i.e. $c_0 = d_0$ and $c_1 = d_1$, which yields

$$T_A(x,t) = c_0 + c_1 t - \frac{c_1 L^2}{2k} + \frac{c_1}{2k} x^2. \qquad (7)$$

The equation (7) provides useful approximate estimate for the time asymptotic temperature distribution inside of the strip/plate, which is being cooled at rate c_1 at both ends.

We also wish to estimate the temperature distribution during the transition from the initial temperature distribution to the time-asymptotic solution (7), and especially the time required for the temperature distribution converge to the time-asymptotic solution. We assume that $T_0(x)$ is the initial temperature distribution inside of the strip/plate, with top and bottom of the strip/plate at temperature c_0, $T_0(-L) = c_0 = T_0(L)$. We denote the $T(x,t)$ as the full time-dependent solution to the differential equation (2). In (Carslaw and Jaeger 1989), time-dependent solutions to this type of problems are given. The solution can be written in the following form

$$T(x,t) - c_0 = u(x,t) + w(x,t), \qquad (8)$$

where the functions $u(x,t)$ and $w(x,t)$ are such that

$$u(x,t) = 0, \text{ when } x = -L \text{ and } x = L,$$

$$u(x,t) = T_0(x) - c_0, \text{ when } t = 0$$

and

$$w(x,t) = c_1 t, \text{ when } x = -L \text{ and } x = L,$$

$$w(x,t) = 0 \text{ when } t = 0.$$

The functions $u(x,t)$ and $w(x,t)$ are then given by (9) and (10) (Carslaw and Jaeger 1989)

$$u(x,t) = \frac{1}{L}\sum_{n=1}^{n=\infty} B_n \sin\left(\frac{n\pi(x+L)}{2L}\right)\exp\left(-\frac{kn^2\pi^2}{4L^2}t\right), \quad (9)$$

where

$$B_n = \int_{-L}^{L} (T_0(x)-c_0)\sin\left(\frac{n\pi(x+L)}{2L}\right)dx$$

and

$$w(x,t) = c_1 t + \frac{c_1(x^2-L^2)}{2k} + \frac{16c_1L^2}{k\pi^3}\sum_{n=0}^{n=\infty}\frac{(-1)^n}{(2n+1)^3}\cos\left(\frac{(2n+1)\pi x}{2L}\right)\exp\left(-\frac{k(2n+1)^2\pi^2}{4L^2}t\right). \quad (10)$$

The time dependence of (9) and (10) shows that the terms in the solution converge towards the time-asymptotic solution (7), proportionally to $\exp(-t/\tau_n)$, where the time constant τ_n is given by

$$\tau_n = \frac{4L^2}{kn^2\pi^2}, \quad (11)$$

where $n=1$ corresponds to the slowest time convergence, and can be used to evaluate the time required for the initial temperature distribution to converge towards the time asymptotic temperature distribution given by (7).

3 Tools

3.1 Heat conduction tool for the hot rolled strips

In order to be able to estimate the time dependent temperature distribution inside of a steel strip, a numerical heat conduction model has been earlier developed (Pyykkönen et al. 2010; Pyykkönen et al. 2012a; Pyykkönen et al. 2012b; Pyykkönen et al. 2013).

Temperature variations within a steel strip cause the phase transformations to start and proceed at different rates. This can cause inhomogenieties in the final microstructure, which affect the flatness and mechanical properties especially for thicker strips or plates.

A web based graphical user interface was developed to interact with the numerical code for the purpose that the development and operation engineers would be able to benefit from the simulation results easily. The graphical user interface is connected to a server side software (see Figure 1) that has two main functions: reading factory data for desired product and start numerical heat conduction simulation corresponding to the industrial strip parameters (strip thickness, water usage at different points along the cooling line etc.).

Using the GUI, the user can modify the actual values that had been used in the real cooling of the strip. The parameters that are used in the detailed cooling model are: the cooling start temperature and the strip speed, cooling water usage at different parts of the cooling line. The model outputs the temperature in the middle of the strip, at quarter thickness and on the top and bottom surfaces as well as at user-defined depth. The output of the simulation is displayed on the GUI as shown in Figure 2.

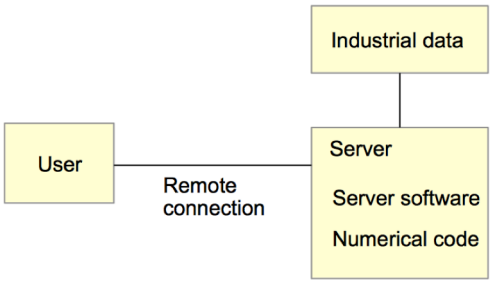


Figure 1. Web browser is used to access the GUI provided by the server software. The server software reads in industrial data, which is used as input for the numerical calculations. The user can modify the input for experimenting how different parameters affect the strip temperatures.

The parameters are passed from the GUI to the numerical heat conduction and phase transformation model as arguments after the executable name in Linux operating system. Following this approach, no swap files need to be generated when starting the computations and further interaction is relatively easy to implement in the codes by adding more arguments to the passing and receiving software. The numerical software produces output files that are then read in and presented by the GUI to the user.

While developing the GUI several project meetings were held where the production and development engineers could suggest modifications to the GUI. This ensured that the functionality provided by the GUI would have the relevant features for the users.

3.2 Phase transformation calculation along user defined cooling path

A phase transformation model, which is currently under active development, is coupled to a graphical user interface, which allows user to define time dependent cooling path for the steel. Such model is useful in finding suitable cooling path that leads to desired final phase composition and mechanical properties of the product. The phase transformation model is fitted for given steel composition and thermo-mechanical treatment in order to obtain as good estimate for the actual transformation behavior of the steel as possible. With numerical experiments using different cooling paths, the time required and cost caused by real experiments can be minimized. The GUI can enable the development engineers to apply even

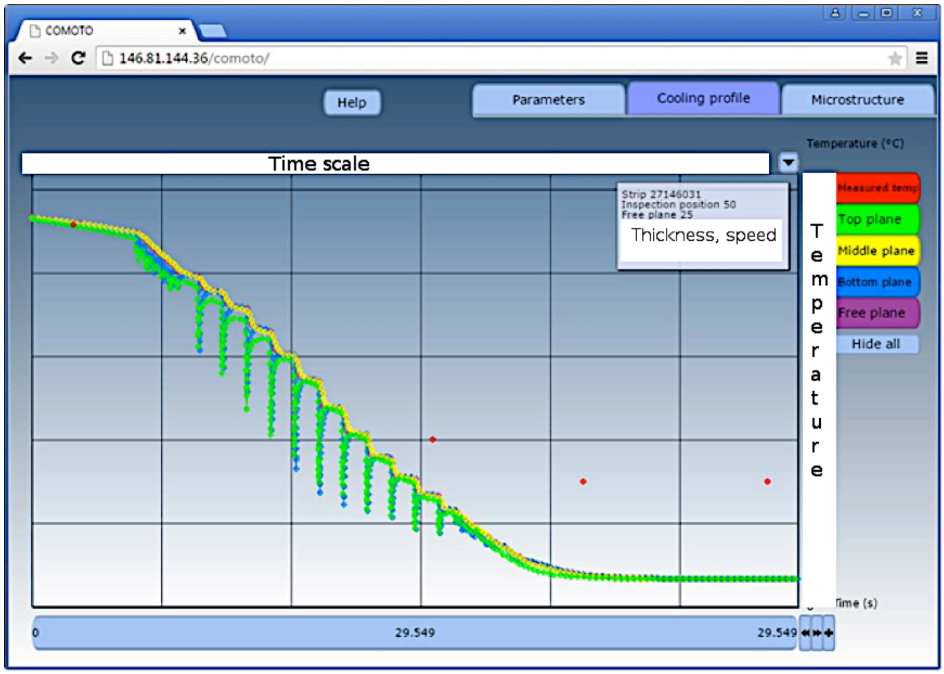


Figure 2. Output from the heat conduction simulation program is visualized with the GUI. The view provides information on the temperature distribution as a function of time.

complicated theoretical transformation model easily. This also benefits the development of the model, as experimental data can be easily compared against the model results.

The graphical user interface is shown in Figure 3. It consists of a diagram with time on the x-axis and temperature on the y-axis. Isothermal transformation start curves of ferrite and bainite are plotted in the diagram to show quickly the transformation start during isothermal holding at different temperatures. The user can define the cooling path by selecting points with a mouse on the diagram. While constructing the cooling path, the user can at any point choose calculate how large fraction of austenite has transformed in to ferrite, bainite, pearlite or martensite.

This GUI was developed with Python programming language with using the modules *tkinter* and *subprocess* (python documentation). The *tkinter* module provides simple syntax for several useful commands, such as drop menu, dialog for saving and loading results, buttons, and canvas for drawing. Simple graphical user interface can be generated quickly, and the modularity of the software provides possibility to further development and combining it with larger projects. Because the syntax of Python language is minimalistic, it provides possibility for rapid implementation.

The phase transformation model on the other hand is developed with Fortran 90 programming language, which enables fast numerical calculations and the use of several optimized numerical libraries. In this case the data is passed from the GUI to the transformation model as a single array in a swap file. The chemical composition is read from an input file. The transformation model outputs the results to a file that is in turn read by the Python code. While the combination of the ease of programming with the Python language and the speed of compiled Fortran code utilizes the main advantages of the both languages and is well suited for small scale projects, such as the one presented here, it may be more difficult to implement in more complicated projects that require more interaction between the GUI and the numerical back-end software. An article where the details of the renewed phase transformation model are described is under preparation and is submitted for publication in 2017 by the first author of this article.

3.3 Calculation of water usage required to cool the mid-depth of a steel strip to a desired temperature

Although the water usage, which is required to cool a steel strip to desired temperatures along the cooling line, could in principle be found by experimenting with different amounts of water and applying the computational method described in section 3.1, this would be cumbersome and slow, since the detailed heat conduction calculation takes 1-5 minutes to complete, depending on the strip thickness. In order to obtain a quick estimate we have developed an approximate method, which calculates how much cooling water is required for cooling the strip at desired temperatures along the cooling line. The model was parameterized using the more detailed model described in section 3.1. The details of this approximate model are presented elsewhere (Paananen 2015, Pohjonen 2016) but here we describe the operation and implementation of the GUI.

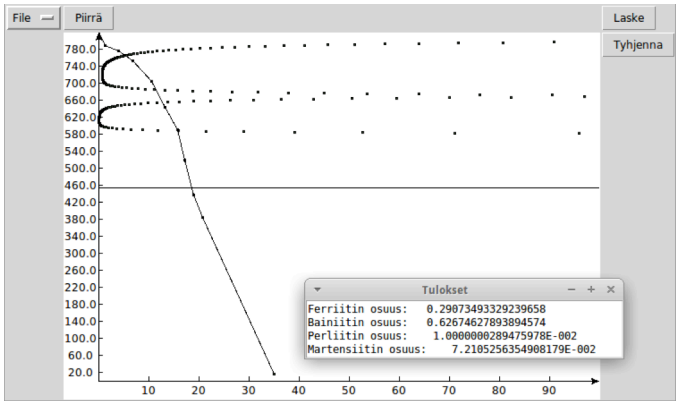


Figure 3. GUI for calculating phase transformations along user-defined cooling path. The user selects the path by clicking on the window and the phases formed can be calculated at any point along the path.

The GUI is shown in Figure 4. The temperature in the middle of the strip is shown in the graph window. The sliders are used to change the desired temperature. User can also load data from the more detailed model in order to compare this quick and approximate model to a more detailed data, calculated by the method described in section 3.1. The user can change the speed and thickness of the strip. Once the waters have been determined, the detailed model can be used to confirm the results and for fine-tuning, if necessary.

In addition to its main goal of determining the cooling water required to cool the strip to desired temperature, the approximate model can also be used in the same way as the detailed model described in section 3.1 to obtain quick and approximate estimate in few seconds (compared to minutes needed by the detailed model).

The GUI as well as the underlying numerical code was programmed with MATLAB. While the interaction of the code with the GUI is ideal within a single language, the non-compiled code is limited in speed. Although this was not a problem in the case described here, this limitation could be overcome by compiling some part of the code and interacting with it through the MATLAB software. (Matlab documentation)

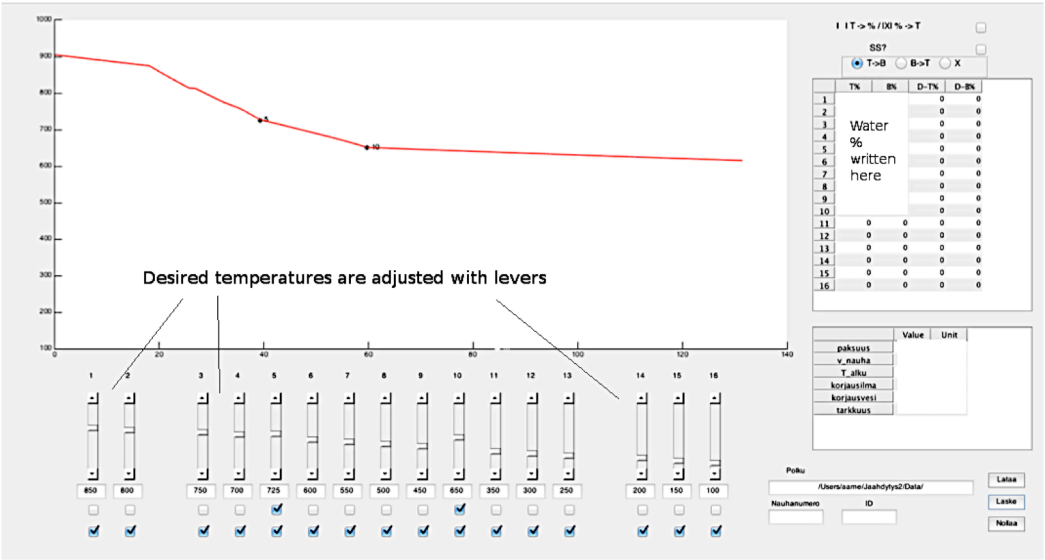


Figure 4. GUI for calculation of required cooling water usage to cool steel strip to desired temperature. The user adjusts the desired cooling water with the levers, defines strip thickness and speed. The required cooling water usage is written on the side panel.

4 Combined use of different tools and analytical approximations

The tools presented in this article provide a way to design a suitable cooling water usage for a steel in order to achieve desired amount of final microstructural constituents (ferrite, bainite, pearlite, martensite). With the use of different tools that have different accuracy, the user can find a quick approximation, which can then be confirmed and fine-tuned with more detailed simulation.

While the heat conduction tool described in section 3.1 can be also fully coupled with the phase transformation tool described in section 3.2, the separate calculation of phase transformations as function of user-defined cooling path is also useful from the design point of view.

The use of different simulation tools and analytical approximation is further illustrated by the following cases

4.1 Planning a cooling water usage to obtain desired amount of ferrite and bainite in the mid-depth

In order to find a suitable cooling path that would produce desired amount of ferrite and bainite in the mid-depth of the strip/plate, we start with trials using the tool described in section 3.2. Assume that we wish to produce 40 % ferrite, 40 % bainite and 20 % martensite for a given steel. Experimenting with different cooling rates, and calculating intermediate results, we can quickly find a suitable cooling rate shown in Figure 5.

After finding the desired cooling path, we wish to estimate the waters required to cool the mid-depth of the strip to this temperature. For this purpose we apply the tool described in section 3.3.

After calculating an estimate for the water usage we can confirm by using the tool described in section 3.1, that the more detailed model gives the same result as the quick estimate.

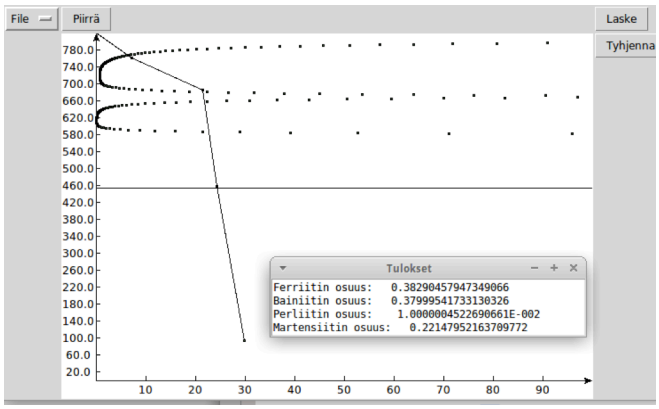


Figure 5. The desired cooling path to achieve roughly 40 % ferrite, 40 % bainite, and 20 % martensite, calculated with the tool described in section 3.2.

4.2 Estimation of temperature distribution for given cooling rate and strip/plate thickness

Since the temperatures in the surfaces can differ significantly from the temperature in the middle of the strip/plate, especially for thicker products, it is useful to understand how the strip thickness and cooling rate affect the temperature distribution within the strip and also to the time needed for the temperature distribution to relax towards the time-asymptotic temperature distribution as described in section 2. It is also useful for the user to be able to calculate quick estimates for the temperature distributions.

The equations presented in section 2 could be easily implemented to a computer algebra system such as the free and open source Maxima CAS (Computer Algebra System) (Maxima webpage) with a graphical user interface wxMaxima (wxMaxima webpage), Maple, Mathematica, or even modern handheld CAS enabled calculators. Sample wxMaxima syntax for the equations presented in this article is available from the first author of this article upon request, and will be made available through the author home page, www.iki.fi/aarne.pohjonen.

We use the following constant values for the thermal conductivity κ = 30 W / (m K) , density ρ = 7400 kg / m^3, and specific heat capacity c = 700 J / (kg K) for the sample calculations. The constant k in the equation (7) is then
$k = \kappa/(\rho c) \approx 5.8 \times 10^{-6}$.

We apply the equation (7) to calculate the time-asymptotic solution for the temperature distribution. For a 10 mm thick strip, L=0.005, subjected to constant cooling rates c_1=–40°C/s on the top and bottom of the strip with the cooling start temperature c_0=1000°C the temperature distribution converges in 3s towards the temperature distribution
$T_A(x, 3s) = 966.3°C - 3.45333\times10^6 x^2$, where the depth coordinate x is in the interval $-L,\ldots,L$ and the mid-depth is at x=0.

The time required to converge towards the solution can be estimated using the equation (11). Since n = 1 gives the slowest convergence, the time constant can be calculated as $\tau_1 = k\pi^2/(4L^2) \approx 1.7$ s. This means that in order for the difference between the asymptotic and the current temperature distribution to diminish everywhere to at most 20% of the original difference, a time $t = -\ln(0.2)\times\tau_1 \approx 2.8$ s is required.

Similar calculations can be applied together with the procedure illustrated with case 4.1 in order to estimate the temperature distributions inside of the strip, when the strip is cooled along the different cooling paths.

Equation (7) shows that the difference between the surface temperature and the mid-depth temperature in the time-asymptotic solution is $c_1L^2/(2k)$, i.e. linear dependence on cooling rate and propotional to L^2. The equations (9) (10) (11) show that the time required for the temperature distribution to converge to the time-asymptotic solution is proportional to L^2. Although the simulation tool 3.1 can be used to calculate the time-dependent temperature distribution for the strip of any thickness and cooling rate, the analytical approximation gives immediately the idea how changing the thickness and cooling rate affect the results.

4.3 Effect of different cooling rates on the top and bottom to the temperature distribution

The flatness of a steel plate is an important property, which should be optimized for the final product. If the temperature distribution of the steel plate is such that other side of the plate is at higher temperature during cooling, the difference in phase transformation start times and thermal expansion can lead to very poor flatness quality.

To see how the different cooling rates on top and bottom of the plate affect the temperature distribution inside of the plate, we apply equation (6) to calculate the time asymptotic temperature distribution when the other side of the plate is cooled with rate $c_1 = -15$°C/s and the other with rate $d_1 = -10$°C/s. We assume 30 mm thick plate, L=0.015. The time asymptotic temperature distribution after 20 s of cooling starting with 1000°C initial temperature, calculated with (6) is shown in Figure 6.

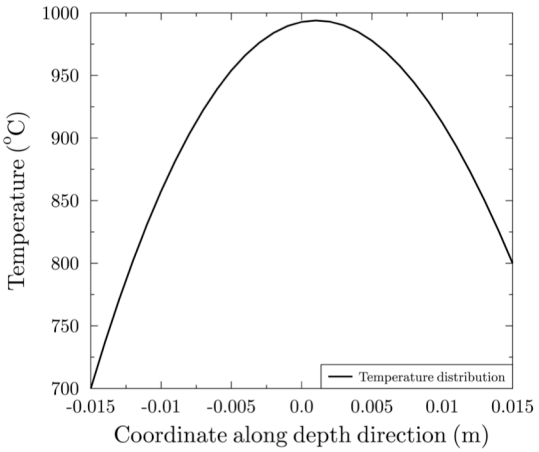


Figure 6. Uneven time-asymptotic temperature distribution $T_C(x,20\text{s})$ caused by difference between the cooling rates on the top and bottom of the strip/plate.

5 Conclusions

We present analytical approximations and simulation tools that can be used for designing cooling path and water usage to cool hot rolled steel strip in order to achieve desired amounts of ferrite, pearlite, bainite and martensite, and understanding how different factors affect the results. The usage of the analytical approximations and the simulation tools is shown by different example cases described in section 4.

Acknowledgements

Useful discussions on the actual dimensions of the steel strips/plates with M.Sc (tech) Olli Leinonen are acknowledged.

References

H.S. Carslaw and J.C. Jaeger. Conduction of Heat in Solids, 2nd edition, Clarendon Press, Oxford, 1989, p.102-105.

Matlab documentation, http://se.mathworks.com/help/matlab/.

Maxima CAS open source computer algebra system, http://maxima.sourceforge.net.

J. Paananen. Laskennallinen työkalu kuumavalssatun teräsnauhan jäähdyttämisen suunnitteluun, Oulun yliopisto, Bachelors Thesis, 2015.

A. Pohjonen, M. Somani, J. Pyykkönen, J. Paananen, and D. Porter. The Onset of the Austenite to Bainite Phase Transformation for Different Cooling Paths and Steel Compositions, *Key Engineering Materials*, 716: 368-375, 2016.

Python documentation, https://www.python.org/doc/.

J. M. Pyykkönen, D. C. Martin, M. C. Somani and P. T. Mäntylä. Thermal behaviour of steel plate during accelerated cooling, *Materials Science Forum*, 638-642: 2706-2711, 2010.

J. Pyykkönen, M. Somani, D. Porter, M. Holappa and T. Tarkka. Modelling of Thermal History and Microstructural Evolution on the Run-out Table of a Hot Strip Mill, *Proceedings from the 6th International Quenching and Distortion Conference*: 817-828, 2012a.

J. Pyykkönen, M. Somani and D. Porter. Experimental and Simulation Studies of Thermal and Microstructure Evolution During Accelerated Cooling of Advanced High Strength Steels, *ROLLING 2013, 9th International ROLLING Conference & 6th European ROLLING Conference*: 1-11, 2013.

J. Pyykkönen, P. Suikkanen, M. Somani and D. Porter. Effect of temperature, strain and interpass time on microstructural evolution during plate rolling, *Journées Annuelles de la SF2M 2012 / SF2M Annual Meeting 2012*, Colloque 1, S1: 17-19, 2012b.

wxMaxima, open source graphical frontend to Maxima CAS, http://andrejv.github.io/wxmaxima/.

Simulation of Horizontal and Vertical Waterflooding in a Homogeneous Reservoir using ECLIPSE

Ambrose A. Ugwu Britt M.E Moldestad

Department of Process, Energy and Environmental Technology, University College of Southeast Norway,
britt.moldestad@usn.no

Abstract

Among the recent deents to improve oil and gas recovery, water injection called waterflooding could be promising. The objective of this work is to ascertain the optimal water injection arrangement between vertical and horizontal waterflooding using ECLIPSE Reservoir simulation software. Within this work, analyses of oil production rate, water cut, reservoir pressure drop, accumulated oil production and recovery factor were made between horizontal and vertical waterflooding in a homogeneous reservoir. Result shows that horizontal waterflooding could be effective if water breakthrough is delayed. The increase in oil recovery achieved through this method varied between 6% and 36% while the delay in breakthrough varied between 459 days and 1362 days. This work also predicts production performance for ten years which would be useful for dynamic optimization of waterflooding. However, reservoir heterogeneity would introduce geological uncertainty, which could bring mismatch between the simulated case and a real case.

Keywords: ECLIPSE, IOR, waterflooding, reservoir

1 Introduction

Waterflooding is a secondary method of oil recovery where water is injected into the reservoir with the aim to increase the pressure and thereby increasing oil production (Binder et al., 1956). Waterflooding was first practiced for pressure maintenance after primary depletion and has since become the most widely adopted IOR technique (Morrow and Buckley, 2011). It is now commonly applied at the beginning of reservoir development(Morrow and Buckley, 2011).

With water injection, the reservoir pressure is sustained and oil is pushed towards the production well. The oil-water front progresses toward the production well until water breaks through into the production stream. With the increasing water production, the oil production rate diminishes, until the time when the recovery is no longer profitable and the production is brought to an end (Van Essen et al, 2006). Up to 35%

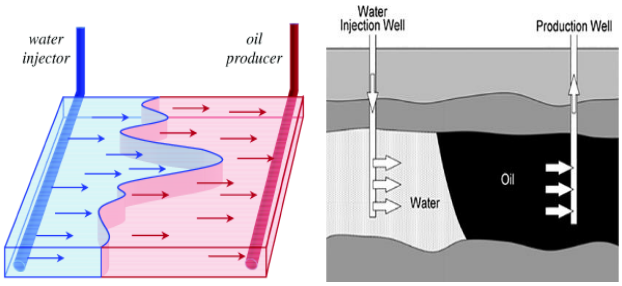


Figure 1. Typical Horizontal and Vertical flooding arrangements (Van Essen et al, 2006): Left (Horizontal waterflooding), Right (Vertical waterflooding).

oil recovery could be achieved economically through waterflooding (Van Essen et al, 2006). Figure1 depicts a typical horizontal and vertical waterflooding arrangement respectively.

Water can be injected through a vertical or a horizontal well. Determining the optimal position and orientation of the wells has a potential high economic impact (Bangerth et al, 2006). One major difference between the horizontal and vertical water injection is the water breakthrough behavior. Asheim studied the optimization of vertical well waterflooding processes with fixed well locations (Zandvliet et al, 2008) while Brouwer and Jansen studied the optimization of waterflooding using a horizontal injection (Brouwer et al, 2001). In both cases, delay in water breakthrough improves production rate. Also from literature, it has been shown that water breakthrough can be delayed by changing the position of the injection well profiles (Brouwer et al, 2001). Studies also revealed that the use of horizontal well, delays the water breakthrough and improves the vertical sweep efficiency (Baker, 1998).

In this paper, computational study of waterflooding in a homogenous reservoir was treated under 6 sections. Sections 1 and 2 deal with the introduction and the theory of waterflooding. Section 3 describes the ECLIPSE mathematical model used in the simulations while Section 4 presents the reservoir model used for the simulations. The simulated results were compared between horizontal and vertical waterflooding in Section 5 with distinct conclusions are in Section 6.

2 Theory

The principal reason for waterflooding is to increase the oil production rate and improve oil recovery. This is achieved through voidage replacement to support the reservoir pressure and sweep or displace oil from the reservoir towards the production well (SPE, 2014). The efficiency of such displacement depends on many factors like oil viscosity, density and rock characteristics. Reservoir screening is necessary for the technical and economic success of waterflooding.

2.1 Residual Oil Saturation

Residual oil saturation and connate water saturation are very important numbers in waterflooding. The connate water saturation is saturation is the lowest water saturation found in situ and determines how much oil is available initially, while the residual oil saturation indicates how much of the original oil in place (OOIP) will remain in the pores after sweeping the reservoir with injected water (SPE, 2014). Equation (1) represents the unit-displacement efficiency with the condition that the oil formation volume factor is the same at the start and the end of the waterflooding (SPE, 2014):

$$E_D = 1 - \frac{S_{orw}}{S_{oi}} \quad (1)$$

$$S_{orw} = 1 - S_{wc} \quad (2)$$

where E_D is the unit displacement efficiency S_{oi} is the initial oil saturation, S_{orw} is the residual oil saturation and S_{wc} is the connate water saturation.

2.2 Wettability

The wettability of a reservoir rock can be defined as the tendency of a fluid to spread on, or to adhere to a solid surface in the presence of another immiscible fluid (Owens and Archer, 1971). In an oil- water system it is a measure of the preference the rock has for either oil or water (Anderson, 1987). Changes in wettability influence the capillary pressure, irreducible water saturation, relative permeability and water flood behavior (Anderson, 1987). Maximum oil production rate by waterflooding is normally achieved at water-wet conditions shortly after water breakthrough (Jadhunandan and Morrow, 1995).

2.3 Capillary Pressure

Capillary pressure is the pressure difference existing across the interface separating two immiscible fluids in porous media. Capillary pressure determines the amount of recoverable oil for waterflooding applications through imbibition process for water wet reservoir (SPE, 2014).

2.4 Relative Permeability

The Relative permeability is the ratio of the effective permeability to the absolute permeability of each phase. It is expressed for a specific saturation of the phases as

$$k_{r,i} = \frac{k_i}{k} \quad (3)$$

where is the phase relative permeability, k is the total effective permeability and is the phase effective permeability.

Relative permeability affects the unit displacement efficiency and how much of the OOIP will be recovered before the waterflooding economic limit is reached. When the interfacial tension between oil and gas phases decreases, the relative permeability values change (Al-Wahaibi et al., 2006), which influences the oil and gas recovery as well as the reservoir pressure. Figure 2 shows the plot of relative permeability curve used for the simulation.

2.5 Mobility

Mobility, λ is described as the ratio between the endpoint effective permeability and the fluid viscosity, μ. It shows how easy the fluid is flowing through a porous medium (Ydstebø, 2013). Mobility ratio, M, plays an important role during waterflooding. It can be defined as the ratio between the mobility of the displacing fluid (water) and the displaced fluid (oil) (Ydstebø, 2013):

$$M = \frac{\lambda_{(displacing)}}{\lambda_{(displaced)}} = \frac{K_{r(displacing)} \cdot \mu_{(displaced)}}{K_{r(displaced)} \cdot \mu_{(displacing)}} \quad (4)$$

where M is the mobility ratio, λ is the mobility, kr is the relative permeability, μ is the viscosity. The subscripts displacing and displaced represent the displacing phase and the displaced phases respectively.

Mobility ratio is considered to be either favorable if the value of (4) is less than or equal to unity or unfavorable if the value is greater than unity (SPE, 2014). Favorable mobility ratio means that the displaced phase (oil) can move more quickly than the displacing phase (water) through the reservoir rock.

3 Computational Model

ECLIPSE Reservoir simulation is a form of numerical modeling used to quantify and interpret physical phenomena with the ability to predict future performance. The process involves dividing the reservoir into several discrete units in three dimensions, and modeling the progression of reservoir and fluid properties through space and time in a series of discrete steps (Schlumberger, 2013). Equations (5-11) are solved for each cell and each time step which are a combination of the material balance equation and Darcy's law (Schlumberger, 2008).

i. Darcy's law (without gravity term) is expressed as

$$q = -\frac{K}{\mu}\nabla P \quad (5)$$

where q is the flux, k is the permeability; μ is the viscosity and is the pressure gradient.

ii. Material Balance is expressed as

$$-\nabla.M = \frac{\partial}{\partial t}(\emptyset\rho) + Q \quad (6)$$

where M is the mobility ratio, ∅ is the porosity, ρ is density and Q is volume flow rate. Here, mass flux is considered as the sum of the accumulation and Injection/Production.

iii. Simulator Flow Equation (with gravity term) is given in (7).

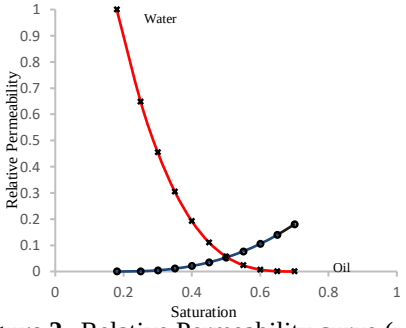


Figure 2. Relative Permeability curve (water-wetted).

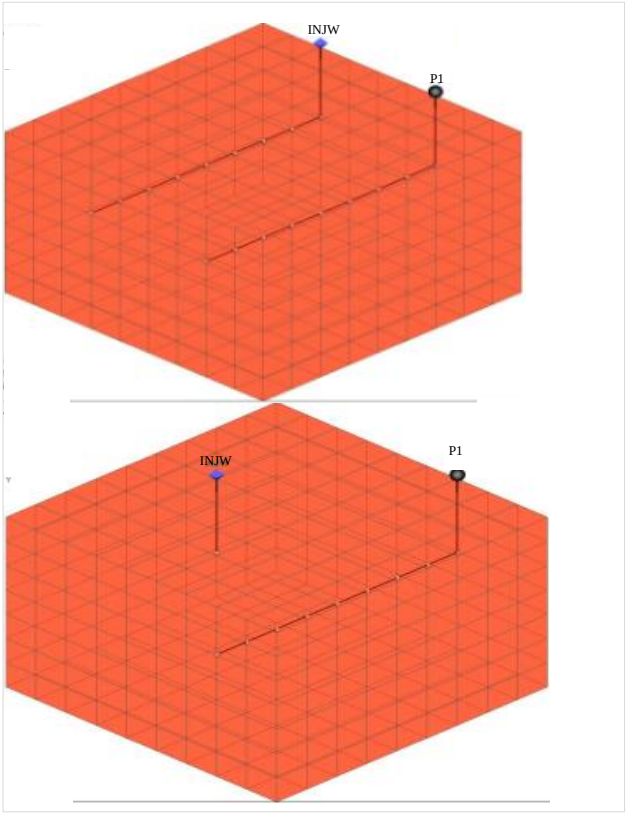


Figure 3. Reservoir Geometry (3D): up(horizontal injection), down(vertical injection).

$$\nabla.[\lambda(\nabla P - \gamma\nabla Z)] = \frac{\partial}{\partial t}\left(\frac{\emptyset}{\beta}\right) + \frac{Q}{\rho} \quad (7)$$

$$\lambda = -\frac{K}{\mu\beta} \quad (8)$$

where M is mobility, t is time, β is momentum transfer coefficient, γ is relative gravity and Z is vertical position.

iv. Well Model is expressed as:

$$q_{p,j} = T_{wj}M_{p,j}\left(P_j - P_w - H_{wj}\right) \quad (9)$$

$$M_{o,j} = \frac{K_{o,j}}{B_{o,j}\cdot\mu_{o,j}} + R_v\frac{K_{g,j}}{B_{g,j}\cdot\mu_{g,j}} \quad (10)$$

$$M_{g,j} = \frac{K_{g,j}}{B_{g,j}\cdot\mu_{g,j}} + R_s\frac{K_{o,j}}{B_{o,j}\cdot\mu_{o,j}} \quad (11)$$

where T is the transmissibility, P is the pressure, H is the pressure head, B is the formation volume factor, R_s is the gas-oil ratio and R_v is the oil-gas ratio. The subscripts p is phase, j is connection, w is well, o is oil and g is gas.

4 ECLIPSE Simulation

Simulations were carried out for 10 years by injecting water at a constant rate through a horizontal and a vertical well respectively. In both cases, water was injected at the same depth as the production well. Also the same lateral distance was maintained between the injection well and the production for both cases. Different simulations were performed by varying injection rate from 200m^3/day to 2,500 m^3/day for each case. A base case without water injection was considered for reference.

4.1 Geometry

Rectangular reservoir geometry was considered with the dimension 900m x 900m x 70m. Figure 3 shows the reservoir geometry for the horizontal and the vertical water injection used in the simulation. The horizontal production (P1) and injection (INJW) wells are 800m long respectively while the length of the vertical injection well (INJW) is 40m.

4.2 Reservoir Conditions

The reservoir is homogeneous and consists of water-wetted rock. Although the reservoir fluid consists of live black oil, gas production was not considered for simplicity. The composition of oil components is assumed to be constant relative to pressure and time. It is also assumed that the reservoir fluid is Newtonian and that Darcy's law applies. Also, the production of light oil in a moderate permeability zone is of interest. The reservoir conditions are summarized in Table 1.

4.3 Initial Conditions

Initially, the reservoir is assumed to be in hydrostatic equilibrium consisting of only oil. It is also desired to have the reservoir pressure above the bubble point to avoid gas production. Initial drawdown pressure of 10bar is also desired. Table 2 shows the initial conditions considered for the simulation.

5 Results and Discussion

In this simulation, analysis of the oil production rate, water cut, reservoir pressure, accumulated oil production and recovery factor were made for the horizontal and vertical waterflooding. A base case without water injection was also considered as reference.

6 Results and Discussion

In this simulation, analysis of the oil production rate, water cut, reservoir pressure, accumulated oil production and recovery factor were made for the horizontal and vertical waterflooding. A base case

Table 1. Reservoir Conditions.

Parameter	Value	Unit
Components	Oil, water, gas	-
Wettability	Water-wetted	-
Porosity	0.25	-
X Permeability	1	Darcy
Y Permeability	1	Darcy
Z Permeability	0.1	Darcy
Rock compressibility	5.0E-5@ 10Bar	/Bar
Oil gravity	35	°Api
Residual oil saturation	0.3	-
Oil viscosity	3 @ 320Bar	cP
Water Density	1000	kg/m^3
Water viscosity	0.5	cP
Connate water saturation	0.2	-
Gas density	1	kg/m^3
Total simulation time	3653	days
No of Grids	567 (9x9x7)	-

Table 2. Initial Conditions

Initial condition	Value	Unit
Reservoir pressure	320	Bar
Bottomhole pressure	310	Bar
Bubble point pressure	182	Bar
Oil saturation	1	-
Water saturation	0	-
Gas saturation	0	-

6.1 Production Rate Trend

Figure 4 shows the oil production rate for horizontal and vertical water injection respectively. The plot shows that horizontal waterflooding maintains higher oil production rate for a longer period until water breaks through. After water breakthrough, the production rate drops more for horizontal waterflooding than the vertical case. This may be attributed to rapid water production in all zones in the horizontal waterflooding case, whereas for the vertical case water breakthrough occurs first in a few zones. The production rate for the base case is very low compared to the cases with waterflooding. This is in agreement that waterflooding improve the oil production rate (Morrow and Buckley, 2011).

6.2 Reservoir Pressure Trend

Figure 5 shows the simulated reservoir pressure trend. For injection rates less than 1500m^3/day, the pressure drop with horizontal injection is between 4% and 6% less than for the vertical case. For injection rates between 1500m^3/day and 2500m^3/day, the pressure drop is between 9% and 14% less with horizontal flooding compared to vertical flooding.

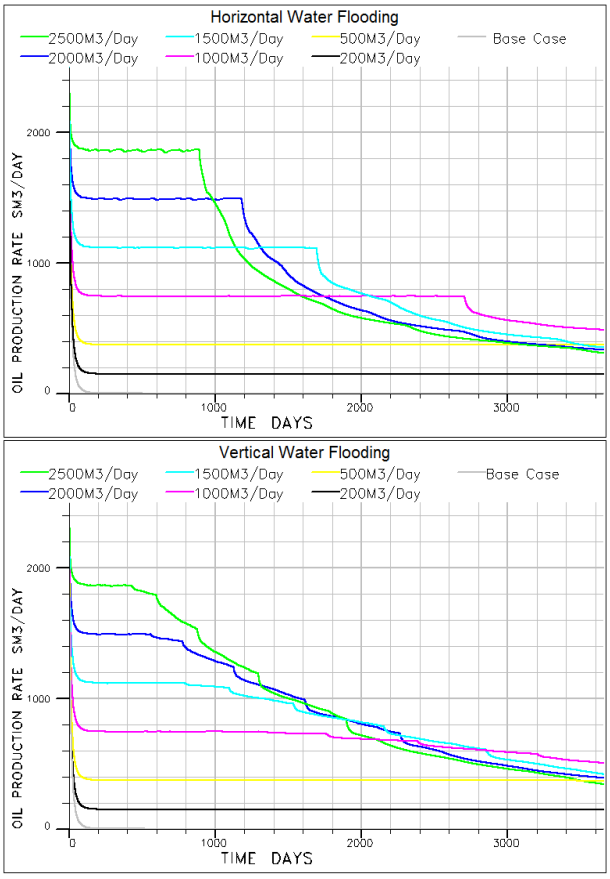


Figure 4. Plot of oil production rate against time: upper plot(horizontal), lower plot(vertical).

6.3 Watercut Trend

The water cut trend is shown in Figure 6. It is observed that water breakthrough is delayed between 459 days and 1362 days with horizontal case compared with the vertical case. Despite of the late water breakthrough, the water cut after 3653 days is higher using horizontal flooding in all the cases.

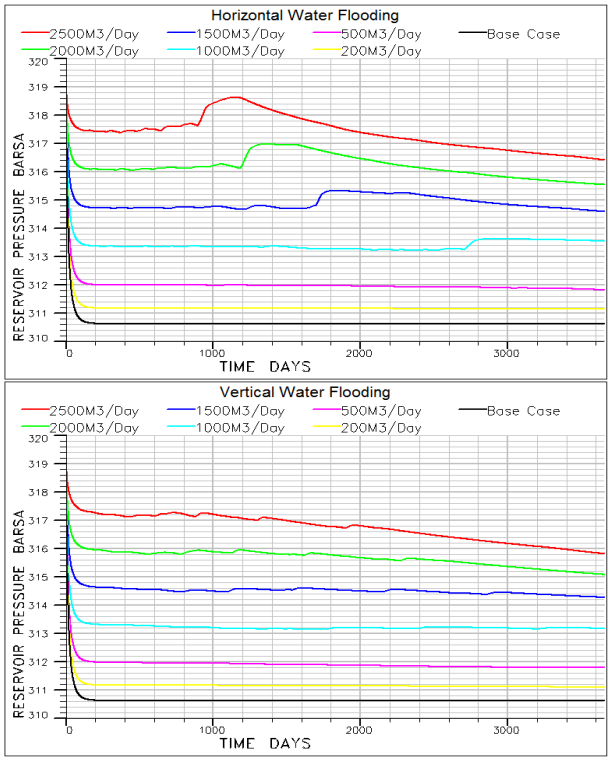


Figure 5. Plot of reservoir pressure against time: upper plot(horizontal), lower plot(vertical).

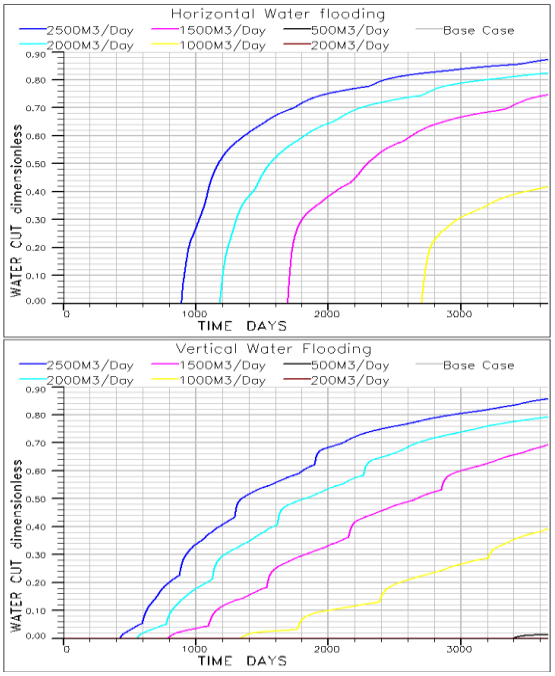


Figure 6. Plot of watercut against time: upper plot(horizontal injection), lower plot(vertical injection).

6.4 Accumulated Oil Production

Figure 7 shows the accumulated oil production trend. The plot shows that the accumulated oil production with horizontal flooding is higher for injection rates less than 1500m^3/day due to lower pressure drop in the reservoir. For injection rates greater than 1500m^3/day, accumulated oil production using horizontal flooding is less than for vertical flooding. This may be attributed to the rapid water production in horizontal flooding as opposed to vertical flooding.

The plot of the recovery factor against injection rate shown in Figure 8 indicates that the recovery factor with horizontal flooding is less than for vertical flooding for injection rates greater than 1500m^3/day. This may be due to rapid water breakthrough.

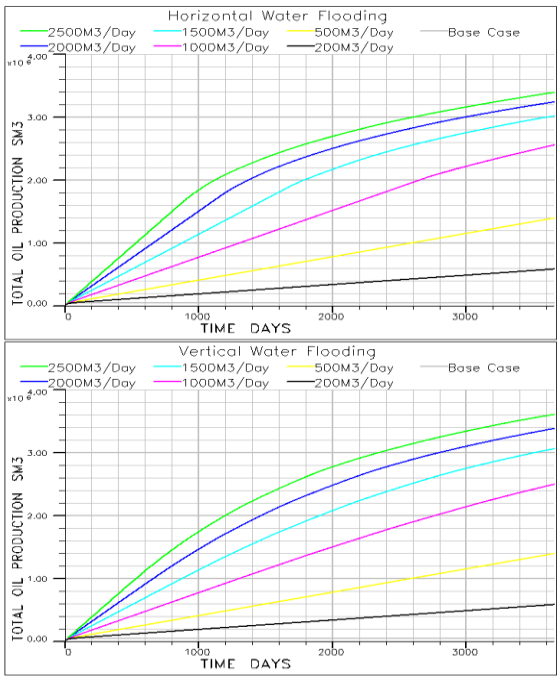


Figure 7. Accumulated oil production against time: upper plot(horizontal), lower plot(vertical).

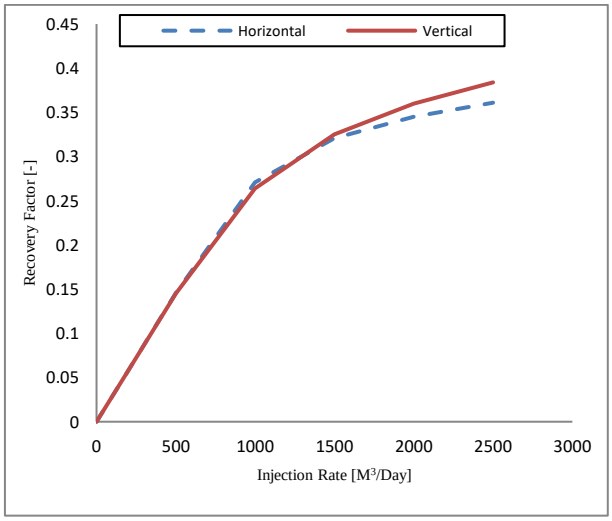


Figure 8. Plot of recovery factor against injection rate.

6.5 Oil Saturation Distribution

The case for water injection at 1500m^3/day is chosen to illustrate how oil saturation is distributed in the reservoir over time for horizontal and vertical water injection respectively. Initially, the oil saturation is 1 for both cases as shown in Figure 3.

Figure 9 shows the oil saturation distribution for horizontal injection after ten years. It can be seen that about 32% oil recovery was achieved through waterflooding. Figure 10 shows the oil saturation distribution for vertical injection after ten years. About 33% oil recovery was achieved through waterflooding.

6.6 Oil-Water Front Progression

The case for water injection at 1500m^3/day is used to illustrate how water displaces oil and sweeps oil towards the production well in the reservoir. Figure 11 shows the plan view of the oil-water front progression for the horizontal and vertical water injection after two years. Figure 10 shows the oil saturation distribution for vertical injection after ten years. About 33% oil recovery was achieved through waterflooding.

6.7 Oil-Water Front Progression

The case for water injection at 1500m^3/day is used to illustrate how water displaces oil and sweeps oil towards the production well in the reservoir. Figure 11 shows the plan view of the oil-water front progression for the horizontal and vertical water injection after two years.

The oil-water front progression after ten years is shown in Figure 12. From the plot, it can be seen that the oil saturation reduced due to more sweep by water injection. In general, result shows that oil-water front progresses laterally for horizontal flooding and radially for vertical flooding.

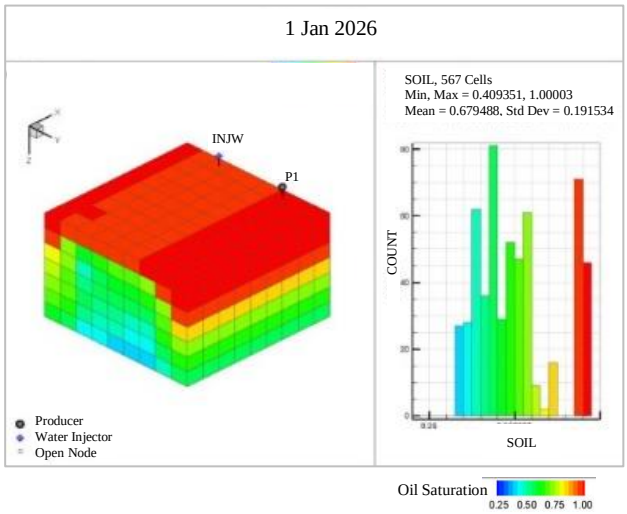


Figure 9. Oil saturation distribution for horizontal injection after 10 years.

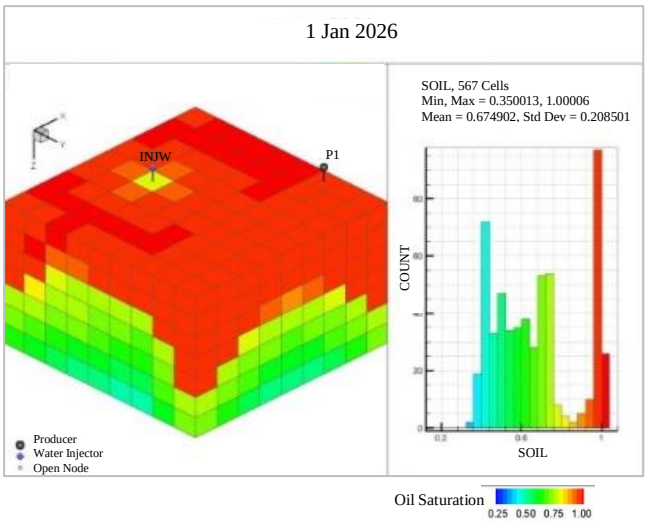


Figure 10. Oil saturation distribution for vertical injection after 10 years.

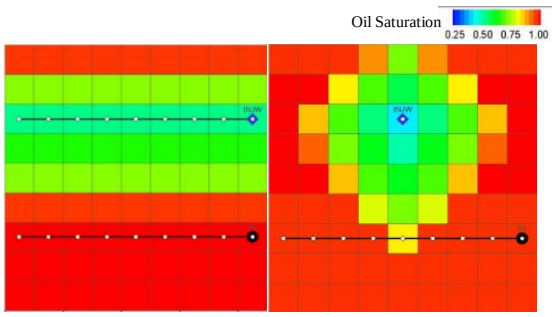


Figure 11. Oil-water front progression after 2 years: left(horizontal injection), right(vertical injection).

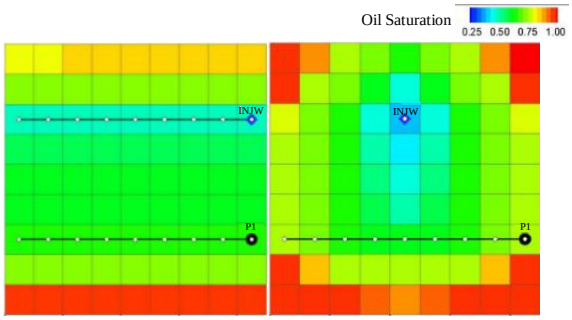


Figure 12. Water front progression after 10 years: left (horizontal injection) right (vertical injection).

7 Conclusions

This paper compares oil production rate, water cut, reservoir pressure drop, accumulated oil production and recovery factor between horizontal and vertical waterflooding in a homogeneous reservoir. The simulation was performed over ten years (3653 days) using ECLIPSE Reservoir simulator.

In all cases, result shows that oil production with water injection is higher compared with the base case.

With this, it would be preferred to apply waterflooding for oil recovery in depleted reservoirs to the use of primary methods. Result also shows that horizontal

waterflooding maintains higher oil production rate for a longer period until water breakthrough. It is also observed that water breakthrough is earlier and water production increases gently with vertical flooding unlike the horizontal case where the water breakthrough comes late but water production increases rapidly with time.

With this, it would be preferred to apply waterflooding for oil recovery in depleted reservoirs to the use of primary methods. Result also shows that horizontal waterflooding maintains higher oil production rate for a longer period until water breakthrough. It is also observed that water breakthrough is earlier and water production increases gently with vertical flooding unlike the horizontal case where the water breakthrough comes late but water production increases rapidly with time.

The pressure drop is higher with vertical flooding in all cases compared with the horizontal flooding. This may be due to higher frictional pressure drop and the effect of gravity. More difference in pressure drop is noticed between horizontal and vertical flooding with increase in injection rate.

Despite higher reservoir pressure and delay in water breakthrough, horizontal flooding accounts for less oil recovery due to rapid water production. With the implementation of inflow control device to reduce water production, oil recovery through horizontal waterflooding would be optimal and more effective than vertical waterflooding.

References

Y. AL-Wahaibi, C. Grattoni and A. Muggeridge, A. Drainage and imbibition relative permeabilities at near miscible conditions. *Journal of Petroleum Science and Engineering*, 53: 239-253, 2006.

W. G. Anderson. Wettability literature survey part 5: the effects of wettability on relative permeability. *Journal of Petroleum Technology*, 39(11): 1,453-451,468, 1987.

R. Baker. Reservoir management for waterfloods-Part II. *Journal of Canadian Petroleum Technology*, 37(1): 300-325, 1998.

W. Bangerth, H. Klie, M. Wheeler, P. Stoffa and M. Sen. Optimization Algorithms for The Reservoir Oil Well Placement Problem. *Computational Geosciences*, 10(3): 303-319, 2006.

J. G. G. Binder, R. C. West and K. H. Andresen. *Water flooding secondary recovery method*. Google Patents, 1956.

D. Brouwer, J. Jansen, S. Van Der Starre, C. Van Kruijsdijk and C. Berentsen. Recovery Increase through Water Flooding with Smart Well Technology. In *Proceedings of the SPE European Formation Damage Conference. Society of Petroleum Engineers*, 2001.

P. Jadhunandan and N. R. Morrow. Effect of Wettability on Waterflood Recovery for Crude-Oil/Brine/Rock Systems. *SPE Journal Reservoir Engineering*, 10(1): 40-46, 1995.

N. Morrow and J. Buckley. Improved Oil Recovery by Low-Salinity Waterflooding. *Journal of Petroleum Technology*, 63(5): 106-112, 2011.

W. Owens and D. Archer. The effect of rock wettability on oil-water relative permeability relationships. *Journal of Petroleum Technology*, 23(7): 873-878, 1971.

Schlumberger Limited. *ECLIPSE Blackoil Reservoir Simulation*, 2008.

Schlumberger Limited. *ECLIPSE Reservoir Simulation Software - Technical Description*, 2013.

SPE (Society of Petroleum Engineers). *Microscopic Efficiency of Waterflooding.* Available via http://petrowiki.org/Waterflooding [Accessed March 5, 2016].

G. Van Essen, M. Zandvliet, P. Van Den Hof, O. Bosgra and J. Jansen. Robust optimization of oil reservoir flooding. In *Proceedings of the IEEE International Conference on Control Applications*, 699-704, 2006.

T. YDSTEBØ. *Enhanced Oil Recovery by CO_2 and CO_2-Foam in Fractured Carbonates.* The University of Bergen, 2013.

M. Zandvliet, M. Handels, G. Van Essen, R. Brouwer and J. D. Jansen. Adjoint-based well-placement optimization under production constraints. *SPE Journal*, 13(4): 392-399, 2008.

Simulator Coupling for Network Fault Injection Testing

Emilia Cioroaica Thomas Kuhn

Embedded Software Engineering, Fraunhofer IESE, Germany,
{Emilia.Cioroaica,Thomas.Kuhn}@iese.fraunhofer.de

Abstract

System architectures of embedded systems are undergoing major changes. Embedded systems are becoming cyber-physical systems (CPS) with open interfaces and resulting distributed control loops. This calls for new testing approaches that enable early evaluation of system and safety concepts, and support the evaluation of system designs before they are implemented. Simulation is a common technology that supports the testing of embedded systems, but existing simulators are focused and specialized. A single simulator often does not support all models needed to provide a valid testing environment for system designs. In this paper, we describe our framework for the coupling of communication simulators to enable virtual testing and safeguarding of embedded system designs. The integration of network simulation models and fault injectors enables testing of safety concepts. The applicability of our approach is illustrated in the context of a case study based on a vehicle system design realized as contract work.

Keywords: virtual testing, simulation, simulator coupling, communicating systems, embedded systems, fault injection

1 Introduction

Enabling the development of safety concepts for open systems is one of the most pressing challenges in embedded systems development. Formerly self-contained and isolated systems are being equipped with open interfaces to enable communication. Closed safety-relevant control loops open up and integrate remote devices via wireless networks. This requires the development of new safety concepts that include detection and handling of potential transmission errors. The development of safety concepts becomes even more challenging when wireless links are used, because these are much less predictable than wired links.

Development and testing of communication systems has always been supported by simulations. They provide a virtual testbed that enables the evaluation of application and protocol aspects. Safety concepts could benefit from this technology as well. However, network simulators are often optimized with respect to performance simulation. The provided protocol and network models accurately resemble timing; errors are, however, only represented by an error flag indicating the presence of a transmission error, and this causes, for example, the dropping of a faulty frame. For the development of safety concepts, a more detailed simulation of faults is necessary. The loss of a frame, for example, is not significant from the viewpoint of a safety engineer. Much more significant are flipped bits, which lead to wrong data being received that is not detected by CRC checksums. Safety concepts need to handle these faults as well. Fault injection testing (Guthoff and Sieh, 1995; Barton et al., 1990; Zussa et al., 2014) is one approach for validating designs with respect to their robustness against faults. However, it only covers functional transmission models that support fault injection, but not performance evaluation.

The development of networked embedded systems requires consideration of both aspects: functional transmission of errors and performance optimization. Safety concepts create additional overhead with their measures: they do not only need to be robust against faults and economical with respect to the additionally used resources, but their failure-handling strategies must not affect other communications beyond a permitted degree either. To support the development of next-generation safety concepts, we have therefore developed a framework for coupling fault injection testing with network simulation. In this publication, we document the integration of different network simulators into our simulation framework FERAL (Framework for Evaluation on Requirements and Architecture Level)(Kuhn et al., 2013), and discuss the integration of fault injection testing for the evaluation of safety concepts for open embedded systems.

The remainder of this paper is structured as follows: Section 2 contains a survey of related work. Section 3 presents the challenges of virtual testing on an open embedded system example. Section 4 documents our framework for the coupling of simulation models for network simulation. Section 5 adds the aspect of fault injection testing. Section 6 presents a case study showing use cases of our virtual testing environment. In Section 7, we draw conclusions and lay out future work.

2 Related Work

Most existing approaches for simulator coupling are tailored couplings between simulators to support testing of systems or to predict the properties of a product under development. This leads to considerable overhead because event detection, simulation accuracy, and the correct coupling of execution models must be considered individually for each coupling. The following references indicate the

concrete necessity for simulator coupling.

The work presented in (Siddique et al., 2007) illustrates the networking domain as another application area of simulator coupling. Through the layered approach used for most communication stacks (Schumacher et al., 2009), this domain is predestined for the integration of simulators. By coupling simulators for link layer protocols and network layers, the authors of (Siddique et al., 2007) build an infrastructure for locating interactions between effects on both layers. The work described in (Schumacher et al., 2009) applies the simulator coupling approach to the automotive domain to simulate the impact of Car-to-X systems. This requires coupling of the OMNeT++ simulator, which provides a network simulation, and the road traffic simulator SUMO. The coupling has to integrate the position data for each vehicle as created by SUMO with the OMNeT++ simulator, so that wireless networking characteristics and the impact of car movements are correctly simulated, including the effects of the Car-to-X application under evaluation.

PicSim (Björkbom et al., 2011) is a tool for the investigation of networked and wireless control systems where two tools, Simulink and ns-2, run on two different PCs, which are connected via LAN. One of the PCs runs a Simulink model in a MATLAB version either for Windows or for Linux. The second PC runs the network simulation and needs to have a Linux operating system. The two simulators are then started at the same time and are synchronized to share their results. In this way, two important parts of the system - functional and network behavior - can be brought together and interdependencies can be tested. The network behavior focuses on wireless communication.

The approach described in (Zhang et al., 2013) couples the notations/tools CarSim, SystemC, and C-Code to simulate and test the behavior of a cyber-physical system. Such a CPS is made up of three parts: a physical layer, a network/platform layer, and a software layer. CarSim is used to simulate the physical layer, while SystemC handles the network/platform layer and the software layer is described by C code. The framework is enhanced by a tool for model-based design to allow the creation of rapid prototypes. The authors of (Eyisi et al., 2012) present the Networked Control Systems Wind Tunnel (NCSWT), an integrated modeling and simulation tool for the evaluation of Networked Control Systems (NCS). NCSWT integrates Matlab/Simulink and the network simulator ns-2 for modeling and simulation of NCS using the High Level Architecture (HLA) (Kuhl et al., 1999) standard.

Existing fault injection approaches focus on the injection of specific faults or classes of faults at the system level (Barton et al., 1990; Guthoff and Sieh, 1995) or at the software level (Duraes and Madeira, 2006) in order to validate the system's functional behavior.

Related work shows that simulator coupling is a commonly applied technique for evaluating communicating systems. Performance evaluation is a proven technique for predicting network performance. Safety-relevant systems additionally need to be evaluated with respect to their resilience. Consequently, fault injection techniques should be additionally used to evaluate network performance, additional overhead due to safety measures, and the impact of faults in one integrated simulation.

3 Virtual Testing of Open embedded Systems

Figure 1 illustrates the functional structure of a remotely controlled lift system, which will be used as an explanatory example for this paper. This system enables remote control of a hydraulic lift with a smartphone. The development of such a system starts with requirements and high-level models for functional behavior and safety concepts. Simulator coupling, as documented in (Kuhn et al., 2013), enables the coupling of these models into one integrated simulation, and therefore the rapid evaluation of concepts and early feedback to developers.

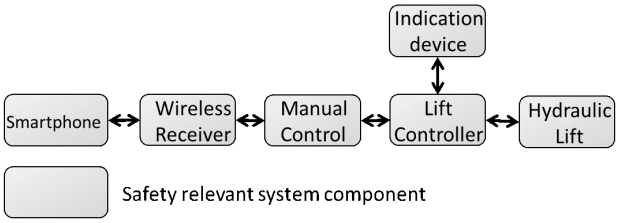


Figure 1. Example Remote Lift Controller System.

For the evaluation of the safety concepts of the example system, a functional structure as shown in 1 is not sufficient. Networks are an integral part of this system, and they affect its safe operation. Safety concepts need to handle network errors reliably. Figure 3 illustrates a revised version of 1 that includes network simulation for the evaluation of safety concepts. Simulation of wireless (LAN) networks is provided by the ns-3 network simulator 0; other simulators are integrated to simulate CAN bus networks and generic wired links. The coupled simulation models split the system into three semantic domains that execute discrete event and discrete time models.

Fault injection testing enables test-based evaluation of safety concepts. To support the evaluation of safety concepts for networked systems, fault injection testing needs to be supported for all network simulation components.

In the context of the example system shown in Figure 3, this includes the ns-3 network simulator, the CAN bus simulation, and the tailored models wired links.

To assure safety, the system behavior needs to be tested in holistic simulations. Faults are injected at the level of simulation components, while behavior needs to be observed both at the component and at the system level. To enable fault injection testing for system designs with different levels of abstraction, fault injection needs to be decoupled from the network simulation.

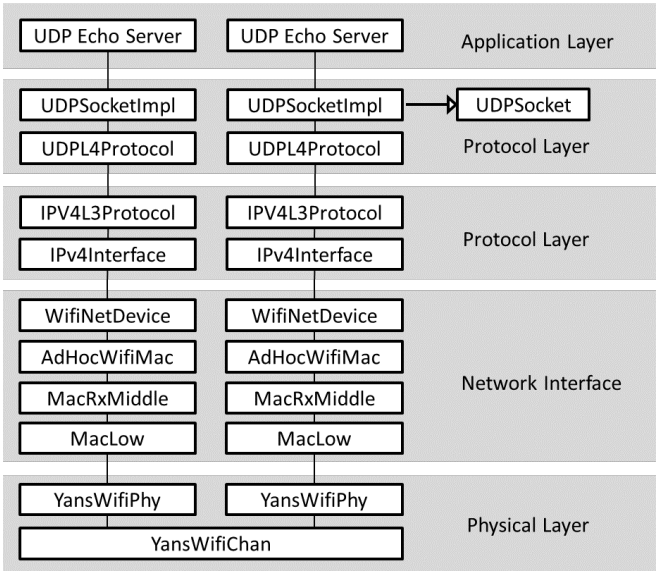


Figure 2. Internal structure of an ns-3 Wi-Fi simulation.

4 Simulator Coupling for Network Simulation

Network simulators offer their simulation models with proprietary interfaces. Development of simulated protocol stacks is possible by combining these interfaces. Figure 4 illustrates the structure of a CAN bus simulation. It consists of the CAN medium, which simulates the CAN bus medium and signal propagation, the CAN bus controllers (CANCtrl), and interface components (CanIF). Interface components aggregate raw data into CAN frames and add offsets, if necessary. On top of the CAN interface components, applications and higher-level protocols like ISOBUS may be implemented.

Developing a generic approach for the coupling of network simulation models and fault injection testing requires an understanding of the inner structure of network simulators, and development of a high-level architecture that encapsulates the components that are specific for each network simulation. Figure 2 and Figure 5 illustrate the architecture of simulated communication stacks for the ns-3 and OMNeT++ simulators.

The example structure shown in Figure 2 illustrates the component instances that create a simulated Wi-Fi network in the ns-3 simulator. Four types of layers can be identified: The Application layer consists of the simulated application behavior. The Protocol layers implement the UDP and IP protocols. The Network Interface layer simulates the medium access control of the Wi-Fi network; its simulation is split into different components. The Physical layer is simulated by components that simulate physical layer encoding and the physical channel. Since ns-3 is a discrete event simulator, frame propagation is based on discrete events as well. Every frame is marked by an event that indicates the transmission of its first bit as well as the frame length. When receiving a physical transmission, every frame is represented by the time that its first bit is received and by the time when its last bit is received. This information is used together with the received signal strength, which is calculated by the propagation model of the physical channel model for the simulation of frame collisions. Upper layers represent frames transmitted and received by one event only. The internal structure of an OMNeT++ simulation is similar. Figure 5 illustrates the component instances that create an Ethernet simulation. It consists of the same basic layers as the ns-3 network simulation. And similar to ns-3, the simulation of the physical channel (EtherBus or YansWiFiChan) is instantiated once; the other components in the Physical layer, the Network Interface layer, and the Protocol layers are instantiated once per simulated network node. Components on the application level may be instantiated multiple times.

The coupling of simulators should yield a protocol stack structure as shown in Figure 4. Applications and application-level protocols should be able to connect to simulated networks. The type of simulated network should be easily replaceable, enabling the simulation of a scenario with both an idealistic network for functional evaluation and its evaluation with a realistic network simulation for performance evaluation.

Figure 6 illustrates our common structure for the integration of network simulation models and functional models. It also shows the types of layers from integrated network simulators that are encapsulated and integrated as simulation components (Kuhn et al., 2013). White-box network simulation models that define all relevant components like queue, medium access control, and application interfaces can be integrated in the same way as black-box components that hide their internal structure. This enables both the rapid integration of existing simulators with low effort as a black-box simulation and the more effort-consuming integration of simulators as white-box components. White-box components require the development of more and additional components, but also enable much finer-grained customization of simulated networks.The component named Application Interface realizes the interface between higher-level protocols and applications, and the network simulation. Its structure is documented in Figure 7.

Applications create PDUs that transmit serialized information through communication networks. In addition to the transmitted data, addressing information might be provided to the simulation components. This addressing information may contain, for example, a CAN bus message ID, or a UDP address including a receiver port number. Adapters may use and change this information to simulate mappings that are realized by the protocol under development. As shown in Figure 7, PDUs additionally contain a list of key/value pairs that store data that is specific to the simulation models and a list of failure modes for this frame.

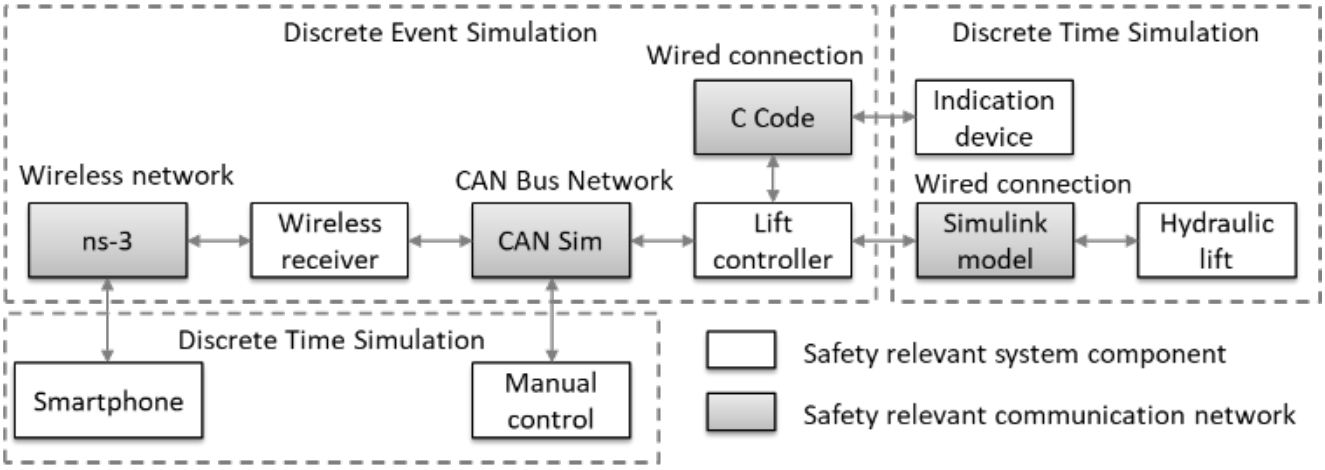


Figure 3. Component types and models of computation of the system example.

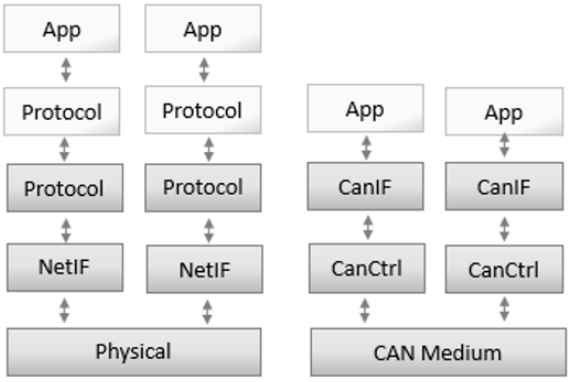


Figure 4. Abstract structure of network simulator.

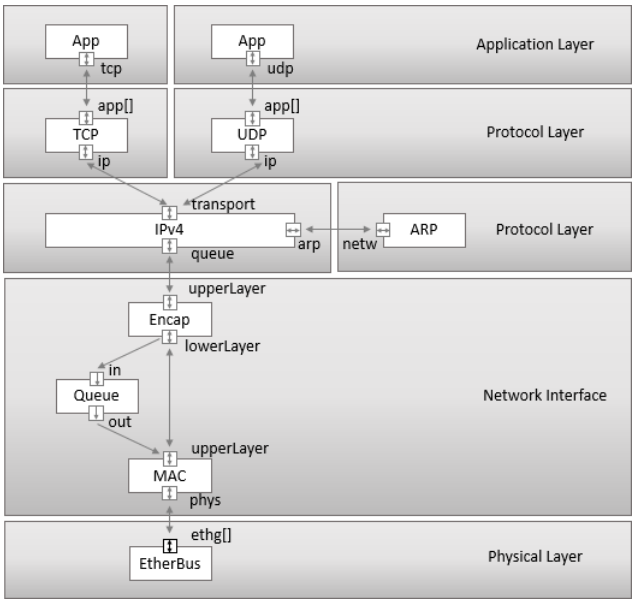


Figure 5. OMNeT++ detailed structure.

5 Network Fault Injection Testing

The integration of network simulators as simulation components enables accurate simulation of networked embedded system components. FERAL enables the rapid development of simulation components and their coupling. Fault injection testing, however, requires a generic approach that is independent of the simulation components that provide the network simulation. Figure 8 lists the relevant failure modes, which were derived from ISO 26262 (ISO 2011) for communication in road vehicles.

While all communication networks basically introduce the possibility of all types of communication failures, the concrete probability of a particular failure depends on the type of communication system to be used. Tailored fault detection and containment mechanisms need to reduce the probability and/or the impact of faults and at the same time conserve resources of the communication system. They add, for example, additional checksums and drop faulty frames instead of passing them to higher protocol layers, preventing the processing of corrupted data by safety-relevant applications.

To effectively simulate failure propagation in embedded systems, the simulation of faults happening and their consequences need to be split from each other. A possible fault on a communication medium is the flipping of multiple bits. A consequence of this fault could be either the dropping of the frame if the bit flip is detected, or a wrong value that is passed to the application.

Our fault injection framework realizes this separation by defining explicit fault injectors and fault processors. Fault injectors create faults in the system. Fault processors simulate consequences of faults. Both fault injectors and fault processors may operate statistically in performance evaluation scenarios to determine performance impacts, or deterministically to evaluate the suitability of a safety concept for the virtual certification of a system.

Figure 9 illustrates fault injection testing with one black-box network simulation component. Fault injectors and fault processors are placed in the communication paths of the network simulation. Both fault injectors and fault processors are added to inbound commu-

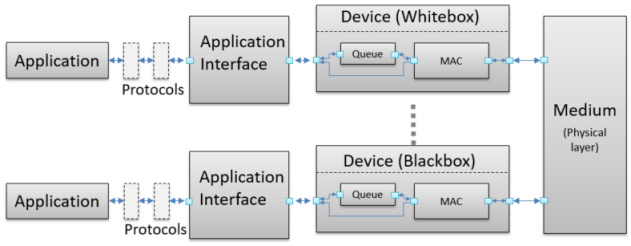


Figure 6. Common structure for integrating network simulation models.

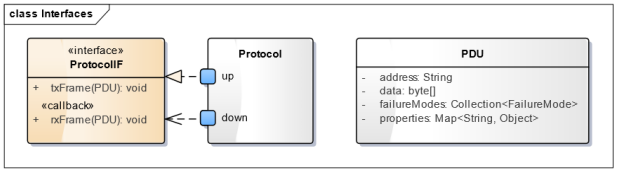


Figure 7. Interface for applications and protocols and PDU structure.

nication paths, whereas only fault processors are added to outbound communication paths. Fault processors in inbound communication paths therefore can control whether adapters will process the faults for the native simulation model or not. Faults that are created by fault injector components are added to the failure-modes list of transmitted PDUs (cf. Figure 7). Fault processors scan the list of faults for each PDU and modify the PDU based on their implementation.

Fault processors and fault injectors enable fault injection testing independent of network simulation components, and consequently their independent development. Therefore, it is best if network simulation models do not realize fault processing at all, but leave the simulation of faults to explicit fault injector and fault processor components.

Figure 10 illustrates fault injection testing with our common structure for the coupling of functional and network simulation models. Fault injectors on medium inputs inject faults that affect all receivers, e.g. interferences from other transmissions or a broken cable. Injected faults are stored in the failure-modes field of transmitted PDUs, as shown in Figure 7. Fault processors on medium in-

Defined failures in E2E communication (according to ISO 26262)
- Repetition of information
- Loss of information
- Delay of information
- Insertion of information
- Masquerading of information
- Incorrect addressing of information
- Incorrect sequence of information
- Corruption of information
- Asymmetric information sent from sender to multiple receivers
- Information from one sender received by subset of receivers only
- Blocking access to communication channel

Figure 8. Failures in end-to-end communication.

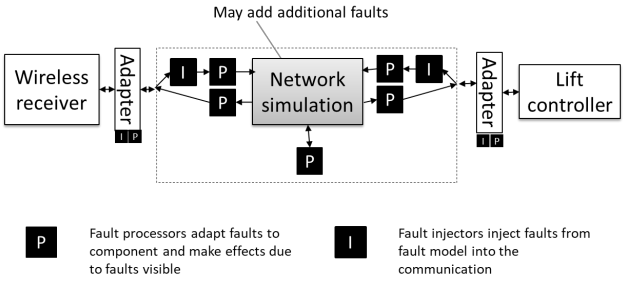


Figure 9. Fault injectors and fault processors.

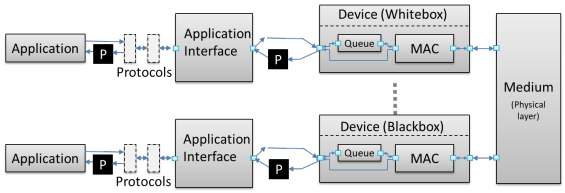


Figure 10. Fault injectors and fault processors.

puts convert injected faults into failures that affect all receivers. Interferences may, for example, lead to shifted bits or dropped frames, or may have no consequence at all if the interference was not sufficiently strong. Fault injectors and processors with medium outputs model effects that only one or a subset of all receivers suffer from. Application-level fault processors enable developers to convert network faults into application-specific faults. For example, they turn undetected bit shifts into application-specific changed inputs. This enables worst-case analysis with critical inputs for application components.

Currently, the fault injectors and processors are available for use at the medium, device, and application level, as in Table 1.

Injectors	Processors
Interference (once)	(Multi) bit change
Interference burts	Frame drop
Signal fading	Frame creation
Cable break	Retransmission
Collision	Delay
Logic error	Sequence change
	Value change

Table 1. Fault Injectors and Processors

Fault injectors and processors support different activation patterns. They can be activated once at a given point in time, in intervals, or sporadically. Fault processors represent both high-level effects caused by protocols, e.g., duplication or the change of frame sequences due to retransmission, and low-level effects like bit changes. At the application level, remaining (uncaught) faults of the interference type may yield a value change that changes frame contents at the logic level.

6 Case Study

The lift remote control example from Section 3 is an anonymized case study based on an industry cooperation project. Horizontal movements of a lift are controlled with a smartphone by at most two operators. Safety measures are applied to ensure that the received user input fully reflects the intention of both users. If contradictory commands are received, movement stops. Periodic heartbeats ensure that communication with both operators is possible. Both the smartphone and the wireless network are considered to be untrusted for system design. Therefore, potential errors in smartphones and networks need to be handled. All safety-relevant processing is performed on the wireless LAN receiver. The case study presented here implements a simulation model that reflects the system model illustrated in Figure 3 with additional fault injection as illustrated in Figure 10. The Wi-Fi network was simulated by the ns-3 simulator; the CAN bus simulation model was developed at Fraunhofer IESE. Corrective actions defined by the safety concept were integrated into the wireless receiver.

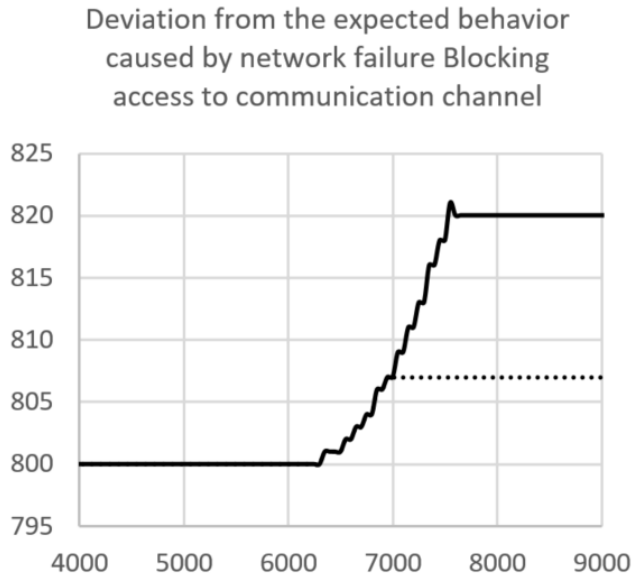


Figure 11. Fault injectors and fault processors.

In this case study, the response of the system to injected faults was evaluated. Once a fault had been injected, it was observed how the system reacted and how well the corrective actions were applied to avoid potential hazards. Figure 11 illustrates one simulated scenario: Due to a simulated network failure, the wireless receiver stops receiving heartbeat messages from the smartphone. As a safety measure, the hydraulic lift stops moving. This is realized by the wireless receiver implementation which, as soon as the heartbeat is not being received anymore, sends stop movement commands to the hydraulic lift. The dotted line represents the user's intended behavior, while the continuous line shows the safe system behavior in the case of the injected fault.

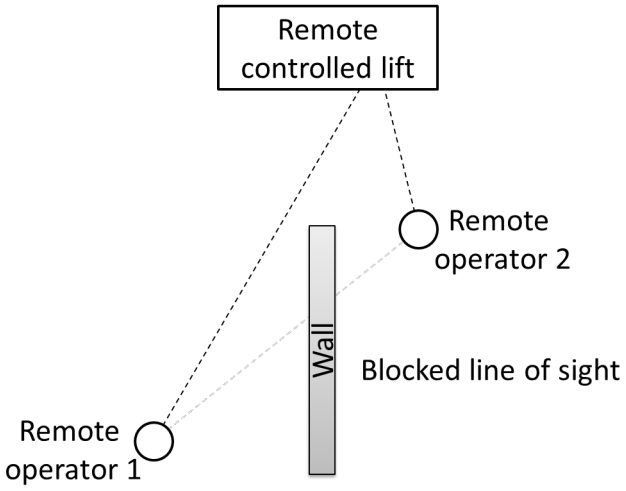


Figure 12. Fault injectors and fault processors.

The second scenario validates another safety function. According to the safety concept, if the two remote operators issue different commands, the lift needs to stop moving immediately. This way, every operator can prevent further movements of the lift if he detects a hazard. Validation of this function requires the integration of a physical network simulation model. As shown in Figure 12, two remote operators are placed at different locations. Due to the topology of the environment, transmissions from the two users to each other are shielded. Therefore, the CSMA mechanism of wireless networks cannot prevent collisions from happening at the wireless receiver. Consequently, control commands collide, which cause the commands with stronger signal strength from operator 2 to be received.

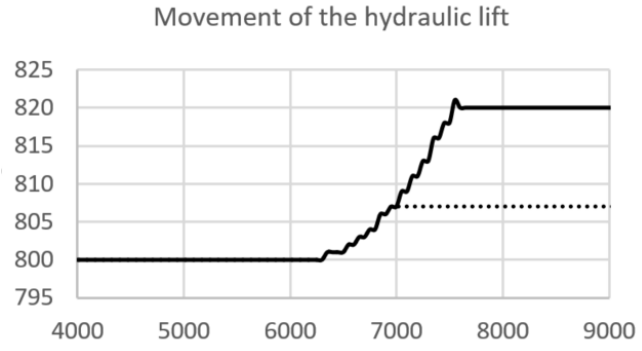


Figure 13. Fault injectors and fault processors.

Frame collisions prevent the more distant operator from communicating with the lift. Due to a sufficient number of heartbeats being properly received, this situation is not detected by the wireless receiver, which is a flaw in the safety concept. As shown in Figure 13, the lift does not stop moving at time 6000, but continues its movement upwards, shown by the solid line. The dotted line represents the expected behavior when two contradictory commands

are issued. The simulation results show that this case was not properly handled by the safety concept

7 Conclusions and future work

In this paper, we presented our work regarding simulator coupling for network and functional simulation, as well as simulator-independent fault injection testing. We analyzed existing simulators and devised a reference structure that represents the most important components of communication stacks. Fault injector and fault processor components enable fault injection testing that is independent of the simulators used. Therefore, failure models are independent of functional models and can be used in both functional and performance evaluation simulations.

This enables the rapid development of virtual communication stacks. Integration of fault injection testing using fault injectors and fault processors enables both statistical injection of faults that resemble the error distribution of a real network.

Future work in this area includes the extension of fault injection testing with platform models. Research needs to be done to evaluate whether memory and processor failures as well as failures in different application models can be integrated into simulations using fault injectors and fault processors in a similar manner as that used for integrating communication failures.

References

James H. Barton, Edward W. Czeck, Zary Z Segall, and Daniel P. Siewiorek. Fault injection experiments using fiat. *IEEE Transactions on Computers*, 39(4):575–582, 1990.

Mikael Björkbom, Shekar Nethi, Lasse M Eriksson, and Riku Jäntti. Wireless control system design and co-simulation. *Control Engineering Practice*, 19(9):1075–1086, 2011.

Joao A Duraes and Henrique S Madeira. Emulation of software faults: A field data study and a practical approach. *IEEE transactions on software engineering*, 32(11):849–867, 2006. doi:10.1109/TSE.2006.113.

Emeka Eyisi, Jia Bai, Derek Riley, Jiannian Weng, Wei Yan, Yuan Xue, Xenofon Koutsoukos, and Janos Sztipanovits. Ncswt: An integrated modeling and simulation tool for networked control systems. *Simulation Modelling Practice and Theory*, 27:90–111, 2012.

Jens Guthoff and Volkmar Sieh. Combining software-implemented and simulation-based fault injection into a single fault injection method. In *Fault-Tolerant Computing, 1995. FTCS-25. Digest of Papers., Twenty-Fifth International Symposium on*, pages 196–206. IEEE, 1995.

Frederick Kuhl, Richard Weatherly, and Judith Dahmann. *Creating computer simulation systems: an introduction to the high level architecture*. Prentice Hall PTR, 1999.

Thomas Kuhn, Thomas Forster, Tobias Braun, and Reinhard Gotzhein. Feral - framework for simulator coupling on requirements and architecture level. In *Formal Methods and Models for Codesign (MEMOCODE), 2013 Eleventh IEEE/ACM International Conference on*, pages 11–22. IEEE, 2013.

Henrik Schumacher, Moritz Schack, and Thomas Kürner. Coupling of simulators for the investigation of car-to-x communication aspects. In *Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific*, pages 58–63. IEEE, 2009.

Mohammad M Siddique, Andreas J Konsgen, and Carmelita Gorg. Vertical coupling between network simulator and ieee802. 11 based simulator. In *Information and Communication Technology, 2007. ICICT'07. International Conference on*, pages 127–130. IEEE, 2007.

Zhenkai Zhang, Joseph Porter, Emeka Eyisi, Gabor Karsai, Xenofon Koutsoukos, and Janos Sztipanovits. Co-simulation framework for design of time-triggered cyber physical systems. In *Proceedings of the ACM/IEEE 4th International Conference on Cyber-Physical Systems*, pages 119–128. ACM, 2013.

Loic Zussa, Jean-Max Dutertre, Jessy Clediere, and Bruno Robisson. Analysis of the fault injection mechanism related to negative and positive power supply glitches using an on-chip voltmeter. In *Hardware-Oriented Security and Trust (HOST), 2014 IEEE International Symposium on*, pages 130–135. IEEE, 2014.

Validation Method for Hardware-in-the-Loop Simulation Models

Tamás Kökényesi István Varjasi

Department of Automation and Applied Informatics, Budapest University of Technology and Economics, Hungary,
{kokenyesi.tamas, varjasi.istvan}@aut.bme.hu

Abstract

The advances in FPGA technology have enabled fast real-time simulation of power converters, filters and loads. HIL (Hardware-in-the-Loop) simulators taking advantage of this technology have revolutionized control hardware and software development for power electronics. Switching frequencies in today's power converters are getting higher and higher, so reducing calculation time steps in HIL simulators is critical, especially if simulating lower power circuits. Faster calculation can be achieved with simpler models or lower resolution. Both possibilities require the validation of the FPGA-synthesizable simulation models to check whether they are correct representations of the simulated main circuit or not. The subject of this paper is a validation method, which treats the simulation error similar as production variance, which can be measured between different instances of the original main circuit.

Keywords: circuit simulation, power circuit modeling, signal resolution, discrete-time systems

1 Introduction

General power converters consist of two main parts: a power stage (main circuit) and a digital controller unit, which is usually realized using a DSP or FPGA. Testing such a controller unit on its original main circuit is expensive and dangerous. That's why offline computer simulation is often used for testing such converters (Rajapakse, 2005; Sybille, 2007). There are very precise models for offline simulation (e.g. PSPICE based simulators), but they can be only used for initial testing of the control algorithms, not the implementation. A low-power model of the main circuit can be built under laboratory conditions, but it will have parameters differing from the ones of the original system.

A very effective way to test controller units' both hardware and software is HIL (Hardware-In-the-Loop) simulation (Kokenyesi, 2013 JEPE). It combines the advantages of other testing methods: low cost like offline computer simulation, complex tests like laboratory testing and realistic conditions like testing on the field. HIL technology also allows the simulation of rare events like failures of certain components which otherwise would be hard to test on an ordinary test bench.

The main concept of using HIL simulation in power electronic systems is that computational models can substitute the high-power parts of the system. These parts can be the power converter itself (Raihan, 2013) or all other power components on both sides of the converter. For example, in the case of a three-phase inverter, models of the motor (Bachir, 2010; Kokenyesi, 2014) or the filter and grid (Kokenyesi, 2013 IYCE). Simulators are connected through real physical interfaces like analog and digital channels to the control boards under development, so they can be tested and validated in their seemingly real environment. A good HIL simulator is completely transparent for the controller unit, so that the controller is unable to distinguish between the simulator and a real system. Therefore HIL simulation can significantly shorten development time and reduce costs (Suto, 2014). HIL simulators are typically realized using FPGA circuits (Cherragui, 2015).

Nowadays and in the near future, switching frequency of power converters is increasing, especially when silicon carbide semiconductor devices are spreading (Biela, 2011). Time constants of these converters are also getting smaller and smaller. To keep the accuracy of HIL simulators acceptable, simulation time steps need to be decreased as well.

To be able to do this, the detail level of HIL simulation models need to be chosen carefully. For example, time constants of snubber circuits are often much lower than the ones of the main power parts, neglecting them can be a significant reduction in computation demand, so the overall result can be improved. Similar situation can occur with parasitic effects; such as serial resistance and saturation of inductors, ESR or voltage dependency of capacitors, semiconductor voltage drops, etc. In some cases, magnetic saturation can be an important effect, which can be the essence of the control loop, so it needs to be simulated properly (Kokenyesi, 2014).

Another possibility to increase the accuracy is to choose a more complex numeric solver for discretization (Kokenyesi, 2013 IYCE). In this case, a good compromise needs to be found between the method's accuracy and the achievable simulation time step. If fixed-point arithmetic is used in the FPGA, the

precision of each variable is also a critical point. How much can they be reduced to make smaller time step possible? What is the minimal required precision?

The proposed validation method is intended to help in these problems. Modeling deficiency, limited resolution or time step produces some deviation in the output signals of the simulated system compared to the real one. This effect is similar to what is caused by standard production variance of the real model's parameters (Kokenyesi, 2014). A properly designed controller unit compensates this deviation (similarly as disturbance signals) and can work with many different instances of the main circuit. The main concept is to treat the modeling error similarly, if it can be compensated, the simulator is passed the validation test.

2 Validation Methods

2.1 Open-Loop Operation

The first approach which comes into view for validation is the open-loop test, its scheme can be seen in Figure 1. In this case, Model A (which is the reference model or circuit) is operated in closed-loop with a properly tuned controller unit and a PWM generator module. Model B (which is under test) is operated in open-loop and its output signals are compared to the ones from Model A. Model B can be a slightly modified (or simplified) version of Model A. It can differ in parameters (like real main circuits), the simulation time step or fixed-point precision. The error in the output signals shows the modeling error. One possible error calculation method is described in (Kokenyesi, 2013 IYCE), which is actually the RMS value of the error. If this is in an acceptable range, Model B considered to be valid and accepted as a HIL simulation model.

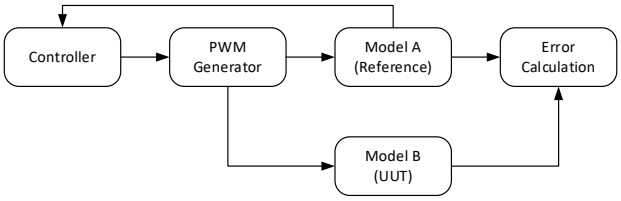


Figure 1. Open-loop validation scheme.

The problem with this method is that the error caused by very small, even negligible model differences is accumulated and it grows continuously as simulation time elapsing. Models with very low damping are especially problematic, e.g. inductors with small serial resistance or resonant circuits with low damping factor. If it was done with two real main circuits, the open-loop operated one would be even damaged after some time.

2.2 Independent Closed-Loop Operation

In this method, the two previous models (A and B) are used again from section 2.1. Both of them are operated with independent controller units with the same structure and tuning, as it can be seen in Figure 2.

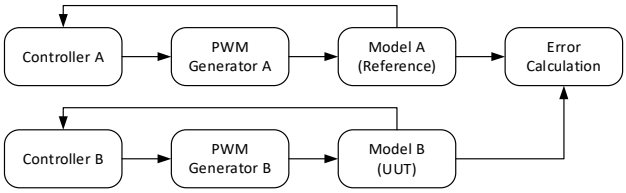


Figure 2. Independent closed-loop validation scheme.

With this approach, accumulated error can be avoided. If a controller is tuned properly for the possible parameter range of the main circuit (or model), it can hide the effect of disturbance signals or model variance and can produce nearly identical output signals. It doesn't mean that all inner variables should be the same. For example, in case of two, slightly different induction motors operated in speed control, the same rotation speed can be achievable with different phase currents. This behavior is one of the main goals in control theory but disadvantageous for this validation, because the simulation error would be eventually hidden.

2.3 Compensated Closed-Loop Operation

The proposed validation method can be seen in Figure 3. Model A is still operated in closed-loop with its properly tuned controller unit. The control signals from the PWM generator are also lead to Model B, with a small intervention based on the output error, which is caused by the difference between the models. The error isn't expected to completely disappear, but it has to be small enough as it would be in the case of two real main circuits. The compensator itself is a special controller unit, which is attempting to reduce the output error to zero. It can modify only the PWM control signals directly, which means inserting switching on or off delays. Otherwise it is similar to the main controller.

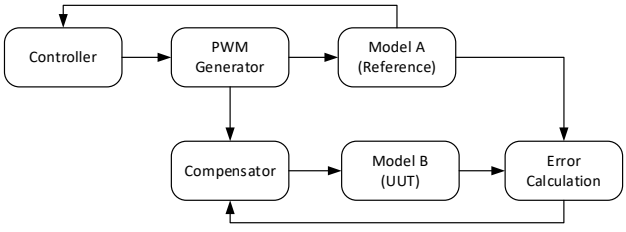


Figure 3. Compensated closed-loop validation scheme.

An important feature of the compensator is that its output is saturated, so the compensable output error is also limited. Switching delays of semiconductors in the real main circuit have a small variance specified in the datasheets. It is a natural difference between the circuit instances, which would be eliminated if they were operated with independent controller units. The controllers would produce slightly different PWM signals, containing the delay variance in this difference. In the validation method, the compensator's output will

contain this variance. If this intervention is saturated to the maximum possible switching delay variance for the semiconductors, the compensator will only be able to eliminate modeling errors, which are less or equivalent to the production variance in effect. If the output error is less than the given tolerance, Model B is accepted as a valid simulation model (Kokenyesi, 2014).

3 Related Work

3.1 Example Circuit

In the following sections, this validation method will be described through a simple example in offline simulation. With two absolutely identical simulation models it is naturally possible to produce the same output signals, so in this case dummy parameter modifications are required to test the validation method.

When choosing the right example, it is important to choose a circuit with very low internal attenuation, which makes it sensitive to proper controller tuning and makes open-loop operation difficult. Taking this into account, a buck converter based battery charger with current control seems to be a good choice. Its schematic can be seen in Figure 4.

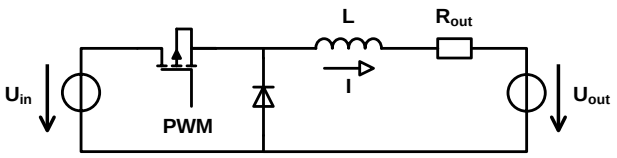


Figure 4. Schematic of the example circuit.

The circuit in Figure 4 contains two DC voltage sources, one for the input DC-link (U_{in}) and another one representing the battery voltage (U_{out}). The semiconductors are considered ideal, only the on and off switching delays (t_{don}, t_{doff}) of the MOSFET are taken into account. Both have a possible minimal and maximal value, which define the acceptable compensator output range. The R_{out} resistance of the L inductor is very small, which makes the attenuation low as required. The control signal is marked as PWM, with f_s switching frequency. The exact parameter values are in Table 1.

Table 1. Parameters of the circuit.

U_{in}	600 V
U_{out}	400 V
I_{ref}	100 A
R_{out}	40 mΩ
L	1.33 mH
f_s	10 kHz
$t_{don,min}$	0.8 µs
$t_{don,max}$	1.2 µs
$t_{doff,min}$	1.8 µs
$t_{doff,max}$	2.2 µs

From these parameters, the switching variance can be calculated:

$$t_{don,diff} = t_{don,max} - t_{don,min} = 0.4\ \mu s, \quad (1)$$

$$t_{doff,diff} = t_{doff,max} - t_{doff,min} = 0.4\ \mu s, \quad (2)$$

$$t_{ddiff} = t_{don,diff} + t_{doff,diff} = 0.8\ \mu s. \quad (3)$$

In the worst case, this t_{ddiff} value is the difference between two instances of the circuit, so it will also be the saturation limit of the compensator.

3.2 Simulation Models

The simulation models of this circuit were built in Matlab/Simulink environment, which is an excellent offline simulation platform, and HDL code generation is also supported for realization of HIL simulators (Suto, 2014). Two different models were made: a floating-point, continuous time model with variable on and off switching delays and a fixed-point discrete time model with variable precision and time step. The two models can be seen in Figure 5 and 6.

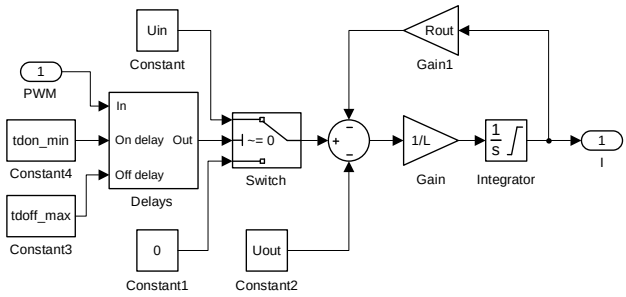


Figure 5. Continuous time model of the example circuit.

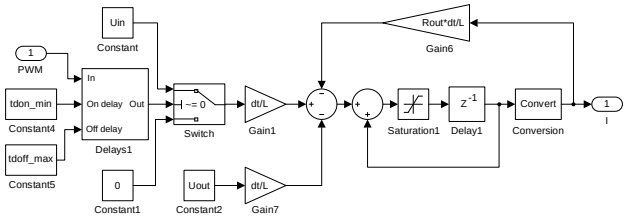


Figure 6. Discrete time model of the example circuit.

Both models contain the same delay generator blocks on the PWM inputs, which allow modifying the switching delays easily. Semiconductors are considered lossless, so a simple switch is used for modeling them. The current integrators are saturated with a lower limit of zero, because of the unidirectional current flowing through the FET and the diode.

In the discrete time model, forward Euler discretization method was used for the integrator (Kokenyesi, 2013 IYCE), where dt is the simulation time step. Because of the fixed-point representation, the integrators' precision are extended to avoid accumulated error (Kokenyesi, 2013 JEPE). The least significant bits are removed with the Convert block after the integrator. Otherwise, the same overall resolution is used for all variables.

For current control, a saturated continuous time PI controller was used, which can be seen in Figure 7. The output saturation ensures that the output voltage doesn't go above the input DC-link voltage and the PWM duty factor remains between 0 and 1. If the saturation is active, there is a difference between this block's input and output signal, which can be used for correction of the integrator and avoid growing its value beyond the limits. It would increase the response time when the sign of the error signal changes next time. When not saturated, this controller operates the same way as any other PI controller, and its tuning can be performed using the traditional methods.

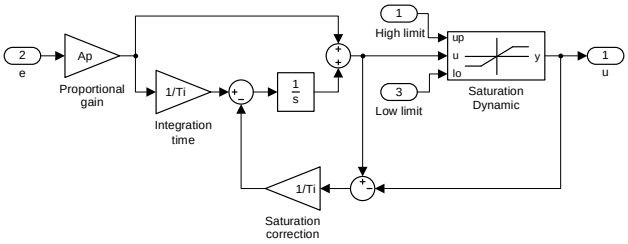


Figure 7. Model of the PI controller.

The compensator itself is also a saturated PI controller with the same structure in Figure 7 and with the same tuning. The difference is only that the saturation limits are calculated from the switching delay variance. The delay differences cause voltage difference, its maximal possible value is the following:

$$U_{diff} = t_{ddiff} f_s U_{in} = 4.8\ \text{V}. \qquad (4)$$

Using this formula, the compensator's voltage output can be converted into delay values, which is used to modify the PWM control signals for the model under validation. In case of positive current error, the compensator produces positive correction output, and the falling edge of the control signal has to be delayed, which causes the FET to switch off later. In case of negative current error, the rising edge and the switching on event has to be delayed, so the current will be reduced.

4 Simulation Results

4.1 Models with Different Switching Delays

First, two continuous time, floating-point models were tested. Differences were only in the switching delays. In the first example, the switching on delay was set to the minimum value in Model A and to the maximum value in Model B. The switching off delays were just the opposite, which means a total t_{ddiff} delay difference. The second example contains two models with switching delay difference $2t_{ddiff}$, which is greater than the allowable maximum.

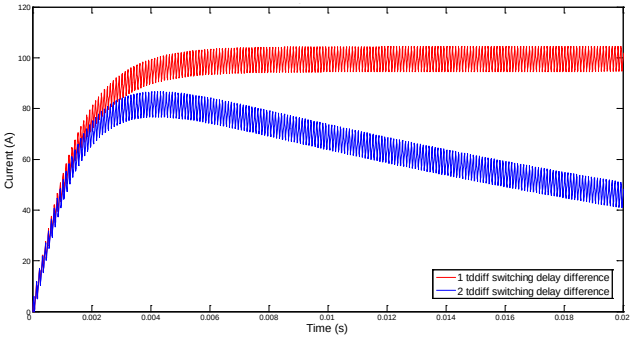


Figure 8. Current signals with different switching delays.

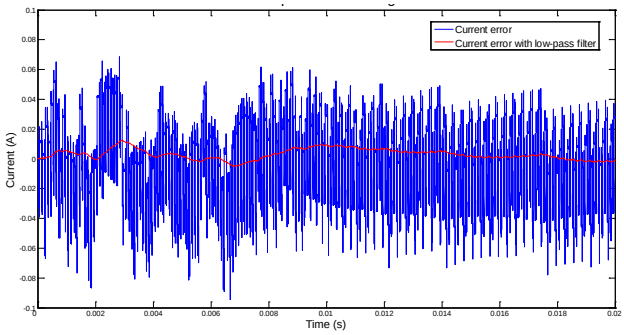


Figure 9. Filtered and non-filtered current error signal (tddiff switching delay).

The simulation results can be seen in Figure 8. The output current signals of Model B in the two simulations are visible. In the case of one t_{ddiff} delay, the current signal can be controlled properly and it is nearly identical to Model A's (which is not in Figure 8). In the other case, the compensator fails to eliminate the current error, which is growing constantly.

It is better to calculate the current error and compare them in different cases than the current signals themselves. In Figure 9 the error is visible in the first case with one t_{ddiff} delay difference. The compensator can only react once every switching period, which causes the switching frequency ripple in the current error. However, it can be stated, that the compensator can keep the current error around zero, additional low-pass filtering of the current error can enhance its visibility as it can be seen in Figure 9. Hereinafter, only the filtered current error signals will be shown.

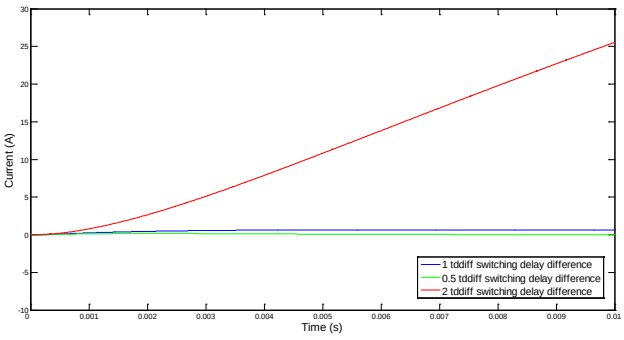


Figure 10. Current error signals with different switching delays.

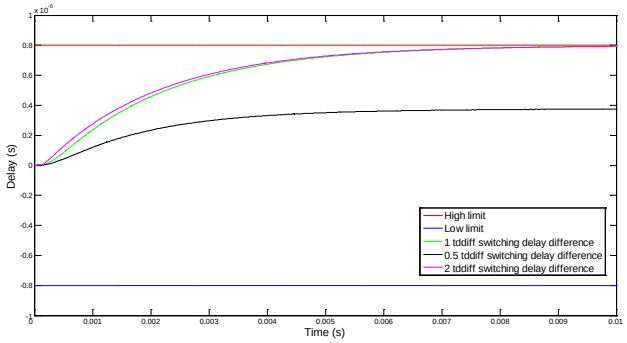


Figure 11. Compensator output signals with different switching delays.

In Figure 10, the filtered current errors are visible in different cases ($t_{ddiff}/2$, t_{ddiff} and $2t_{ddiff}$ delay differences). In the first two example, the current error can be eliminated by the compensator, while in the third one, it is growing. The compensator's output delay is in Figure 11. It is also filtered similarly as the current error. The limits of the compensator's output is also visible. In the case of $t_{ddiff}/2$ delay difference, the intervention in the control signals remains in the allowable range. In the other two cases, it reaches the limit. When the delay is exactly t_{ddiff}, it is just enough to eliminate the error, when it is larger, the error is growing linearly.

These examples were used to test the basic functionality of the compensator and the validation method. It can compensate the original switching delays which correspond to the production variance, so it can be tested on discrete-time models too.

4.2 Discrete-Time Models

When discretizing the model, choosing the simulation time step is critical. This validation method helps to determine the minimum required time step to achieve the required accuracy. It is shown through the following examples.

First, a discrete time model with 100 ns time step was simulated as Model B. Other parameters were the same as the reference continuous time model (Model A) as well as the switching delays. Two additional simulations were run: one with a 2 µs time step and one with a 2.5 µs time step, all other parameters were the same. The general variable precision was 18 bits, uniformly.

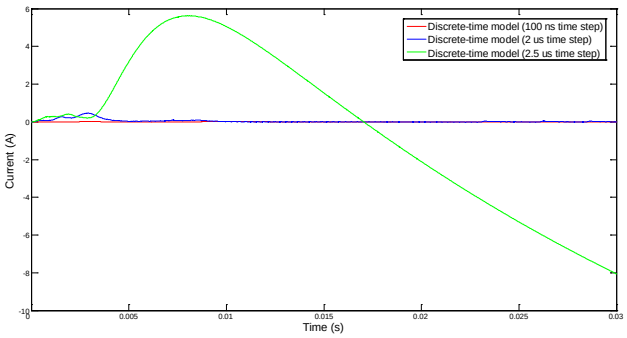


Figure 12. Current error signals with different simulation time steps.

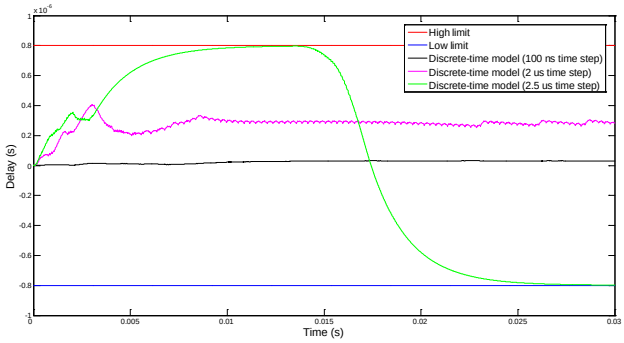


Figure 13. Compensator output signals with different simulation time steps.

The current errors and compensator outputs are in Figure 12 and 13, respectively. In the first two cases, the compensator stays in the normal operating range and the filtered error remains roughly zero. This shows that the model is valid; its simulation error can be compensated as it would be standard production variance.

The last example was simulated with 2.5 µs time step. It is clearly visible that the model is not valid with this time resolution, the current error is growing constantly after an initial transient, while the compensator's output is saturated. It is important to mention that it is not the numerical instability of the forward Euler method (Kokenyesi, 2013 IYCE) which causes the problem, because the system's time constant is 33 ms, which is much higher. The simulation is stable, but not accurate enough.

4.3 Fixed-Point Models

Another important aspect of discretization is to choose an adequate fixed-point representation for the variables. The proposed validation method is also usable to determine the minimum required resolution similarly to the time step.

Three different resolutions were tested as Model B: 18 bits, 9 bits and 8 bits. Like the time step test, all other parameters were the same as the continuous time floating-point Model A. The simulation time step was 100 ns in all cases.

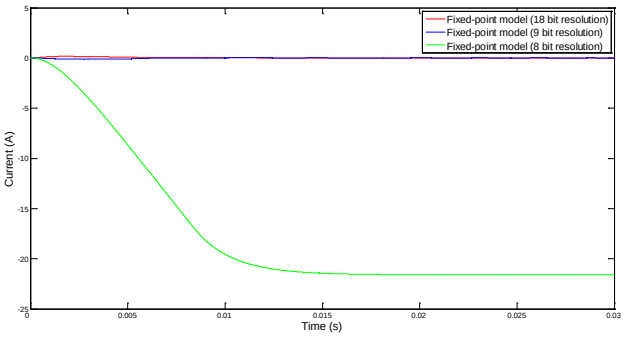


Figure 14. Current error signals with different fixed-point precision.

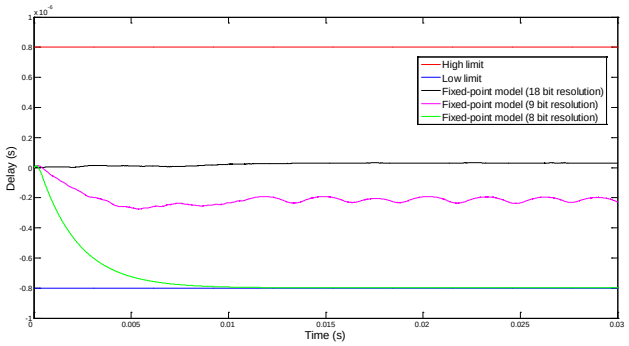


Figure 15. Compensator output signals with different fixed-point precision.

The current errors and compensator outputs are in Figure 14 and 15, respectively. In the case of 18 or 9 bit precision, the compensator eliminates the error successfully with allowable output signals. When using 8 bit precision, the current error stabilizes at -20 A and the compensator is saturated, so 9 bits seems to be the required minimum precision.

5 Conclusions

In this paper, a validation method for HIL simulation models was introduced, treating the simulation errors like the effects of production variance of the main circuit. It considers the simulation model valid, if small intervention in the control signals can compensate the model's error. The limits of the intervention are defined from the production deviation (catalogue data), so there is no essential difference between valid models or real main circuits from the controller unit's aspect. The minimal required simulation time step or fixed-point precision can be determined using this method.

Acknowledgements

This work was performed in the frame of FIEK_16-1-2016-0007 project, implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the FIEK_16 funding scheme.

References

T. O. Bachir, J. P. David, C. Dufour and J. Belanger. Effective FPGA-based Electric Motor Modeling with Floating-Point Cores. *Proceedings IECON'2010, Glendale (USA)*: 829–834, 2010. doi:10.1109/IECON.2010.5675179.

J. Biela, M. Schweizer, S. Waffler and J. W. Kolar. SiC versus Si—Evaluation of Potentials for Performance Improvement of Inverter and DC–DC Converter Systems by SiC Power Semiconductors. *IEEE Transactions on Industrial Electronics*, 58(7):2872–2882, 2011.

H. Cherragui, M. Hilairet and S. Giurgea. Hardware-in-the-loop simulation of a boost converter with the Xilinx System Generator from Matlab/Simulink. *Proceedings IECON'2015, Yokohama (Japan)*: 1837–1842, 2015. doi:10.1109/IECON.2015.7392368.

T. Kokenyesi and I. Varjasi. FPGA-Based Real-Time Simulation of Renewable Energy Source Power Converters. *Journal of Energy and Power Engineering*, 7(1):168–177, 2013.

T. Kokenyesi and I. Varjasi. Comparison of Real-Time Simulation Methods for Power Electronic Applications. *Proceedings IYCE'2013, Siofok (Hungary)*, 2013. doi:10.1109/IYCE.2013.6604137.

T. Kokenyesi. FPGA-based Real-Time Simulation of a PMSM Drive. *Proceedings AACS'2014, Budapest (Hungary)*: 160–171, 2014.

S. R. S. Raihan and N. A. Rahim. Comparative Analysis of Three-Phase AC-DC Converters Using HIL-Simulation. *Journal of Power Electronics*, 13(1):104–112, 2013. doi:10.6113/JPE.2013.13.1.104.

A. D. Rajapakse, A. M. Gole, P. L. Wilson. Electromagnetic transients simulation models for accurate representation of switching losses and thermal performance in power electronic systems. *IEEE Transactions on Power Delivery*, 20(1):319–327, 2005. doi:10.1109/TPWRD.2004.839726.

Z. Suto, T. Debreceni, T. Kokenyesi, A. Futo and I. Varjasi. Matlab/Simulink Generated FPGA Based Real-time HIL Simulator and DSP Controller: A Case Study. *Proceedings ICREPQ'14, Cordoba (Spain)*: 1–16, 2014.

G. Sybille, H. Le-Huy, R. Gagnon and P. Brunelle. Analysis and Implementation of an Interpolation Algorithm for Fixed Time-Step Digital Simulation of PWM Converters. *IEEE International Symposium on Industrial Electronics'2007, Vigo (Spain)*: 793–798, 2007. doi:10.1109/ISIE.2007.4374698.

Embedded Simulations in Real Remote Experiments for ISES e-Laboratory

Michal Gerža[1] František Schauer[1,2] Petr Dostál[1]

[1] Faculty of Applied Informatics, Tomas Bata University in Zlín, Zlín, Czech Republic, gerza@fai.utb.cz
[2] Faculty of Education, Trnava University in Trnava, Trnava, Slovak Republic, fschauer@volny.cz

Abstract

The paper focuses on the design of the module of embedded simulations for real remote Internet School Experimental System (ISES) experiments. The ISES experimental platform is intended for educational purpose laboratories at schools and universities providing computer oriented measuring environment for Engineering students and students of Natural sciences. At present, the ISES remote laboratories do not provide any provisions for concurrent interactive simulations in the form of virtual experiments. This drawback results in lesser attractiveness and understanding of real world phenomena. The designed solution uses the Easy JavaScript Simulations (EJS) environment to calculate the data, using equations provided by physics laws, and the ISES module for the simulated data transfer and visualization.

Keywords: RLMS, ISES, measureserver, phenomena simulation, remote laboratory, real experiment

1 Introduction

Traditional teaching methods for students at schools and universities are outdated and not effective enough. Students often expect faster and more understandable teaching methods in the field of physics, chemistry and electro-engineering, which can help them to better perceive real world phenomena. The problem is also an accessibility of the educational materials, especially for distant students who nowadays prefer studying scientific topics using their computers via the Internet. These problematic points are effectively solved by the remote laboratories (RLs) so called e-laboratories. The RLs built on the ISES platform (Lustig, 2009) have been developed for long time since 2002 by the RL Consortium of three universities (Charles University in Prague, Tomas Bata University in Zlín and Trnava University in Trnava) for educational purposes. The ISES is an advanced tool for real-time operation, data acquisition and processing.

The platform is an open system consisting of the ISES components for a hands-on experimentation called ISES WIN. However, there is also an alternative for using remote experiments (REs), the ISES WEB Control Kit. Both types of the experiments are built as the burst (fast) and normal (slow) to offer students a wider spectrum of knowledge. The initial version of ISES RLs has been developed by Charles University in Prague. When the ISES RL became the time-proven educational tool, it was significantly improved to a higher level tool by Tomas Bata University in Zlín in cooperation with Charles University in Prague. They implemented features for the new user environment, EASY REMOTE-ISES, to build REs by laymen (Krbeček *et al*, 2013).

Let us describe the ISES RL concept. It consists of five units as the HW components (signal convertor, control board, physical modules categorized as meters, sensors and devices), Measureserver, Imageserver, Webserver and Webclient. More details and deployed applications are available in (Zeman, 2011; Zeman, 2012; Krbeček *et al*, 2014; Hamid, Modammed, 2010; Drigas *et al*, 2006). All the ISES RLs were integrated into a new system platform called Remote Laboratory Management System (RLMS) REMLABNET accessed on www.remlabnet.eu. It provides units as services to the ISES RL administrators and clients. The schematic arrangement of all the involved autonomous units and defined communication is presented in Figure 1.

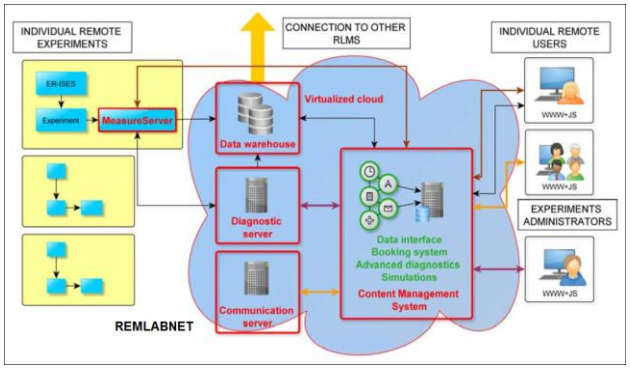


Figure 1. Scheme of the REMLABNET covering ISES remote laboratories with the experiments and clients.

2 State of the art

The ISES RL units internally exploit monitoring, controlling, and communication functions to cooperate with other units to dispatch measured data.

2.1 Physical hardware

There are two concepts in principle implemented for the ISES laboratories - local and remote, built on the same physical HW. The ISES is a modular platform based on three parts. As the first part used, it is the signal convertor installed as the PCI 1202 interface card inside a control computer. Further parts are the control board and the set of sensors for physics, chemistry and electro-engineering. The platform offers a possibility of the data measurement, data visualization and analysis. A complete set of the physical HW is illustrated in Figure 2.

Figure 2. ISES physical hardware including the PCI 1202 interface card and a broad range of the involved meters, sensors and devices.

2.2 Measureserver unit

The Measureserver (MS) unit is a significant software part of the ISES RL concept. It is the processing and communicating server located between the physical HW and remote clients. The MS core is designed as an advanced finite-state machine to setup and process the logical instructions solving prescribed activities. Its functioning is drawn from the concise process script (PSC) file loaded to the MS before its startup.

With respect to the physical HW, the MS in reality communicates with the PCI 1202 interface card. This is the entirely digital process based on the direct reading of data (real values) from particular pins and writing of data to respective pins, which are translated by the signal convertor. These data pins are inputs and outputs located on the control board allowing an access to the particular physical modules like meters, and devices.

Instructions (specific commands), coming from a remote client, are processed by the listening MS. The communication is realized by standard protocols via the Internet. Certain commands go via the MS translator to the REMLABNET where clients can exploit additional services provided like the acquirement of measured data results from previously performed REs stored in the exposed database (warehouse) to analyze phenomena.

All the commands are processed in a deterministic way by two different underlying parsers. The first is called the LR(1) parser that processes commands from the configuration file for the purpose of the graphical user interface settings. This parser is based on static state transition tables (parsing tables), which are able to codify a given language grammar. These parsing tables are parameterized together with a lookahead terminal. This lookahead establishes the maximum tokens, the parser can use to decide, which rule it should use.

The second is the Recursive descent parser (RDP) that processes commands coming from the PSC source to create the defined data structures and logic schemes for the RE. It uses a general form of top-down parsing where backtracking may be involved. The parsing algorithm is based on the walking through a tree.

2.3 Webserver unit

The Webserver unit provides the Nginx services coming into the process when client enters a web page of the ISES RL via the REMLABNET platform. The Nginx is an open source reverse proxy server for HTTP, HTTPS, SMTP, POP3 and IMAP protocols.

2.4 Webclient unit

This unit is a graphical interface provided by web pages via the Internet allowing registered clients to simply control and observe a respective RE (either on the REMLABNET or EU RLMS Go-Lab platforms).

3 Example of real remote experiment and its mathematical model

Simulations play an increasingly important role in the way we teach or do science. This is especially true in education, where computers are being used more and more as a way to make lectures more attractive to students, and to effectively help them achieve a deeper understanding of the subject being taught. This section deals with the design and realization of the phenomena simulations module (PSM) that is able to run concurrently with a respective real RE on the client's web page. The PSM is a part that does not need to be activated every time during the experimentation.

The PSM is designed as the optional autonomous module integrated into the MS structures in cooperation with the Easy Java Simulations (EJS) core to solve the evolution by a preset numerical method. Mathematical expressions, e.g. constants, variables, algebraic equations and ordinary differential equations (ODE), are defined in the PSC. It is notable to mention, the PSC is a textual source having own syntax of programming language similar to the C language. It is parsed by the RDP inside the MS core during the RE

startup. The new objects and additional functions were implemented to identify, initialize and perform the simulation process. Simulated data, produced by the EJS solver, are being continuously transferred on the web page to either charts or tables.

For the purpose of the demonstration how the simulation may be embedded in a RE we use the measurements of the response of the passive parallel RLC circuit to a voltage perturbation in a time domain as shown in Figure 3 showing C capacitor, L inductor (with internal resistance RL) and variable resistors R1D and R2D; ISES voltmeter and amperemeter serve for measuring voltage and current response in the time domain to the perturbation by a unit step voltage, produced by the DC source U(0) and the switch S at the time t = 0. Variable parameters are used two resistive components, artificially introducing the damping. The desired results are all three parameters of the RLC circuit examined. The corresponding initial client's web page with all the preset widgets is illustrated in Figure 6. Let us next describe the problem in a more detailed way.

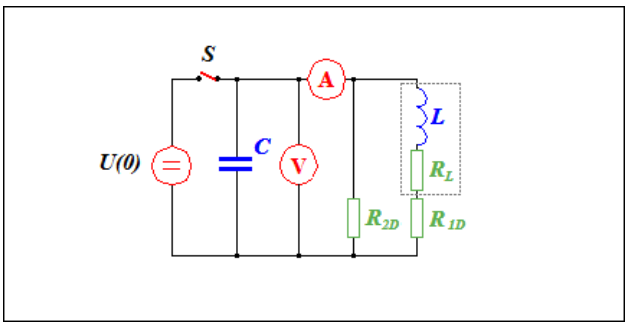


Figure 3. Schematic diagram of the parallel RLC circuit including three resistors.

3.1 Mathematical expressions definition

The mathematical description of the circuit shown in Figure 3 is a form of Kirchhoff's law for a parallel RLC circuit. For the numerical solution of the current *I=I(t)* and voltage *U=U(t)*, both parameters of the circuit - C, L, R_L and damping resistors R_{1D} and R_{2D}, together with the initial conditions should be adjusted and varied. The goal of the using along with the RE is to find the parameters of the circuit by varying both the damping resistors.

There are defined two substitutive resistors induced from Figure 3 to use for further operations.

$$R_1 = R_{1D} + R_L \tag{1}$$

$$R_2 = R_{2D} \tag{2}$$

From Kirchhoff's voltage law (KVL), the following second order differential equation can be constructed

$$\frac{d^2u}{dt^2} + 2b\frac{du}{dt} + \omega_0^2 u = 0 \tag{3}$$

where

$$2b = \left(\frac{1}{R_2 C} + \frac{R_1}{L}\right), [b] = s^{-1}, \tag{4}$$

$$\omega_0^2 = \left(\frac{1}{LC}\right)\left(1 + \frac{R_1}{R_2}\right), \tag{5}$$

$$\omega_1^2 = \omega_0^2 - b^2. \tag{6}$$

The solution, when the essential condition $\omega_0 > b$ is satisfied, follows in the form

$$u(t) = u(0)e^{-bt}\sin(\omega_1 t + \varphi), \tag{7}$$

$$\delta = bT = \ln\frac{u(t)}{u(t+T)} \tag{8}$$

where b is defined as the damping factor and δ is a logarithmic decrement.

Quantity e^{-b} defines how the amplitude relatively makes smaller per unit of time. Equation (7) presents damped oscillations observed in Figure 4, when all the involved elements are set correctly.

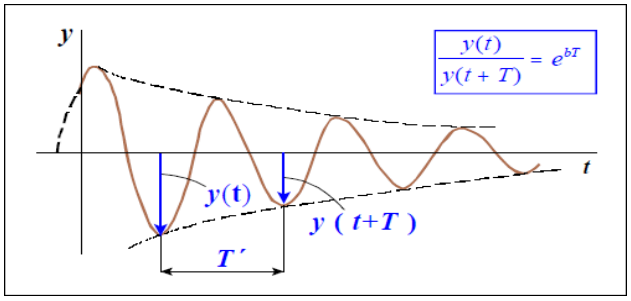


Figure 4. Demonstrative oscillations of the damped parallel RLC circuit.

The simulation process of the parallel RLC circuit with the variable damping is constructed by (3) with the exploitation of (4) and (5), and using original resistors R_{1D} and R_{2D} shown in (1) and (2) as follows.

$$\frac{d^2u}{dt^2} + \left(\frac{1}{R_{2D}C} + \frac{R_{1D}+R_L}{L}\right)\frac{du}{dt} + \left(\frac{1}{LC}\right)\left(1 + \frac{R_{1D}+R_L}{R_{2D}}\right)u = 0 \tag{9}$$

Second order differential equation (9) must be rewritten to a form as shown in Figure 5 to pass it with other parameters to the EJS solver integrated into the PSM. The goal is to obtain and visualize data of the voltage in the given circuit illustrated in Figure 3.

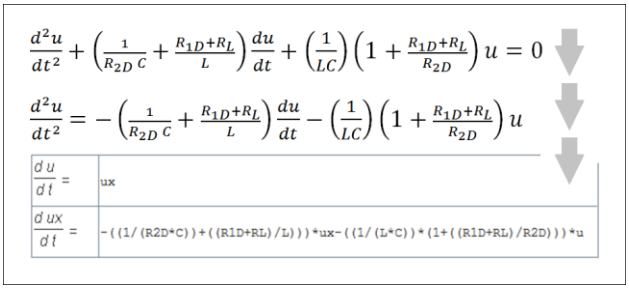


Figure 5. Rewriting (conversion) of the second order ordinary differential equation to a readable form used by the EJS solver.

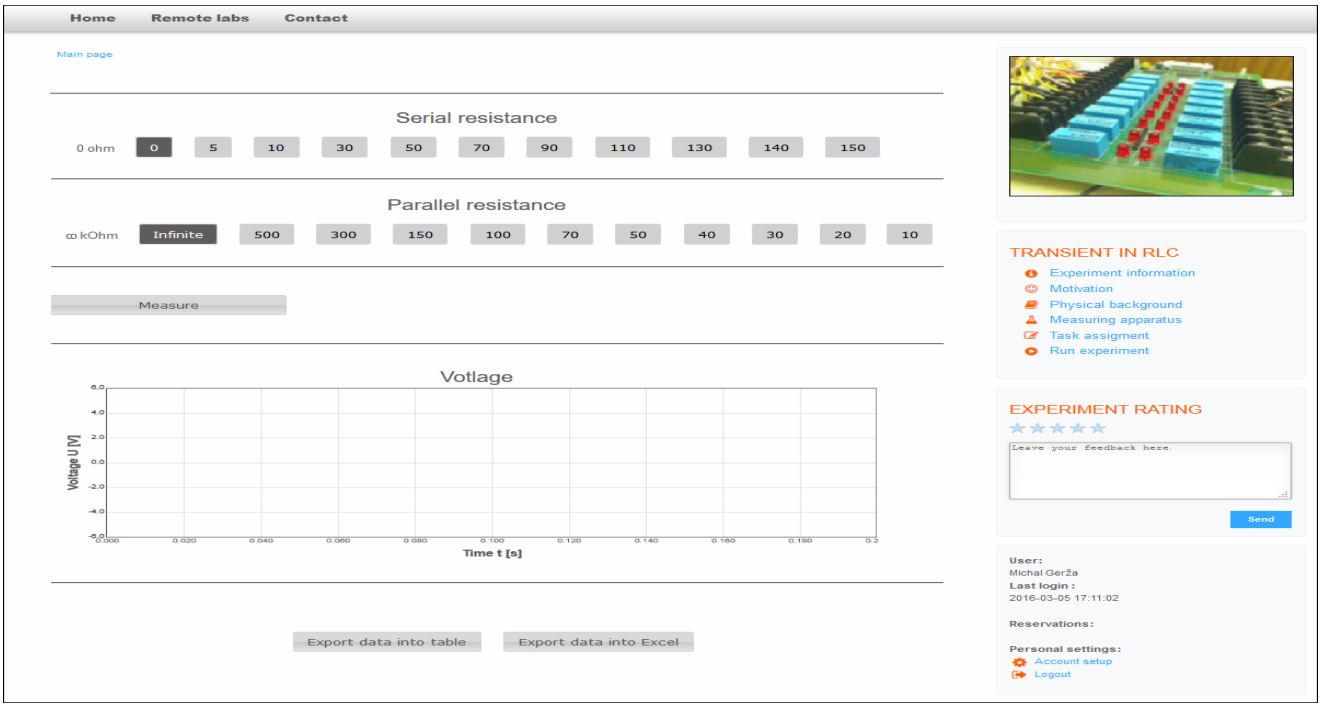


Figure 6. Web page of the ISES remote experiment "Transients in RLC"; the serial resistance R_{1D} and parallel resistance R_{2D} are allowed to set before running the voltage measurement. Resultant voltage progress is displayed in the chart and it can be also exported into two different data tables.

When all the mathematical expressions are properly defined, the EJS parser then generates a respective XML simulation file. In the next phase, the XML is referenced in the PSC control program file where the control process of the simulation is constructed by new commands to set inputs and outputs, including the solver startup when the initialization is complete.

Finally, the PSC is passed to the MS unit that performs all the commands to create and run the real RE together with the concurrent simulation process of the parallel RLC circuit with the variable damping.

3.2 Simulation process script definition

After the differential equation is rewritten into the readable form for EJS, the next step is saving of the transformed form of the simulation assignment into a respective XML file and its full path reference is then inserted into the PSC control file of the RE.

There are generally two alternatives, how to build in the simulation into PSC file. The first one and more complicated, is its direct inclusion into PSC control file, i.e. to code all the definitions manually. The RE designer should individually decide to which steps to connect the simulation. This is a more complicated alternative, necessitating a good knowledge of the programming language, intended just for programmers.

The second alternative, faster and more comfortable, suited to teachers, is the use of EASY REMOTE-ISES providing an intuitive graphical environment for the design and configuration of the RE and its simulation.

The PSM providing the phenomena simulation is an optional feature and must be first declared and enabled in the header of the PSC file. An example, illustrating the coded sequence of functions configuring the concurrent simulation process with the respective variable coefficients (for the resistor, inductor and capacitor), is shown in Figure 7.

3.3 Measureserver core process

The MS unit is an important component in the process chain of completion of all calculations to reach the phenomenon simulation and its time synchronization with the real RE running concurrently.

There are five main activities the MS must pursue to provide all inevitable services as listed below:

1. *Parsing:* The RE control program with the simulation process is parsed by the RDP at its startup from the PSC file.
2. *Controlling:* The installed ISES devices are being controlled in specified time interval or scheme.
3. *Monitoring:* The ISES meters and sensors are being monitored in specified time interval.
4. *Archiving:* The measured data and metadata are being archived to the XML and LOG files and dispatched into the warehouse residing in the REMLABNET to fold and analyze.
5. *Simulating:* The PSM calculates an evolution of the respective phenomenon and generates resultant data for the visualization and analysis.

```
1.  name rlc_circuit

2.  simulation enabled
3.  using simulation_src "RLCCircuit.xml"

4.  EJSSimType simulator

5.  experiment raw
    {
6.      init
        {
7.          simulator = RunSolver();
        }

8.      on_sample card_lib
        {
9.          #operations intended for the real experiment
        }

10.     finalize
        {
11.         simulator.Synchronization();
12.         #finalization of operations for the real experiment
        }
    }
```

```
13. pin_read simulation_voltage
    {
14.     result = simulator.voltage();
    }

15. pin_read simulation_current
    {
16.     result = simulator.current();
    }

17. pin_write simulation_resistor
    {
18.     simulator.R = new_value;
    }

19. pin_write simulation_inductor
    {
20.     simulator.L = new_value;
    }

21. pin_write simulation_capacitor
    {
22.     simulator.C = new_value;
    }
```

Figure 7. Coded sequence of implemented functions of PSM to enable, configure and run the simulation process stored in the PSC file.

The MS activities, mentioned in the five above points, are shown in Figure 8. There are depicted relationships among the PSC control program file, MS unit, ISES physical experiment and REMLABNET

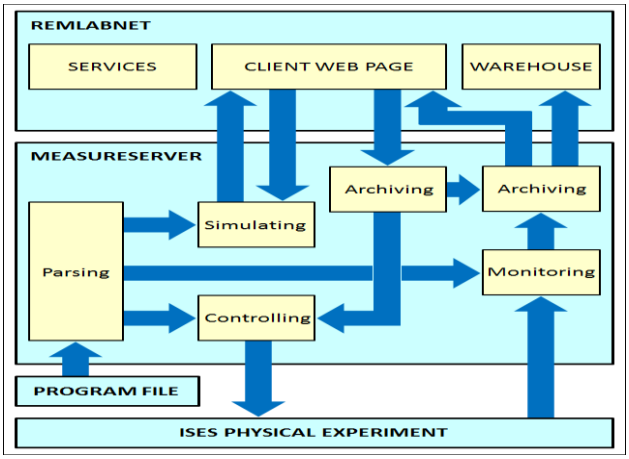


Figure 8. Communication relationships and the activities of the MS unit to realize the real ISES RE and concurrent phenomenon simulation.

3.4 Easy Java Simulations core process

Easy Java Simulations is a software tool designed for the creation of simulations. The EJS is a modeling and authoring tool expressly devoted to science teachers and students. It has been designed to let its user work at a high conceptual level, using a set of simplified tools, and concentrating most of the time on the scientific aspects of the demanded simulation, asking the computer to perform all the other necessary but easily automated tasks.

Nevertheless, the final results, which are generated by the EJS from a user description, can be taken, in terms of efficiency and sophistication, as the implementation of programmer (Esquembre, 2015).

Since the EJS is Open Source under a GNU GPL license, we decided to exploit its optimized solver to perform phenomena simulations integrated into the PSM as a component part of the MS unit.

The EJS solver computes a numerical approximation to the solution of an ODE or, more precisely, of an initial value problem. That is, given the system of ODEs and the state of the system at a given time. The solver is able to approximate the solution of the ODE in a future time. Solver algorithms are all one-step methods, mostly based on Runge-Kutta (RK) schemes, can be explicit or implicit, fixed-step or adaptive, and all use an interpolation to provide a dense output. It means these algorithms produce solution points at any instant of time.

The PSM mostly uses two EJS solvers only. The first solver is the Runge-Kutta 4. It refers to the classical RK method which started it all. It is a fixed step, 4th-order algorithm that works well in most situations. This solver interpolates using one step of the bootstrapping algorithm applied to the Hermite interpolation (this gives order 4 interpolation). The second one is the Cash-Karp 5(4) that is an adaptive 4th order method. It is based on two embedded RK schemes of order 5 and 4. We use a local extrapolation

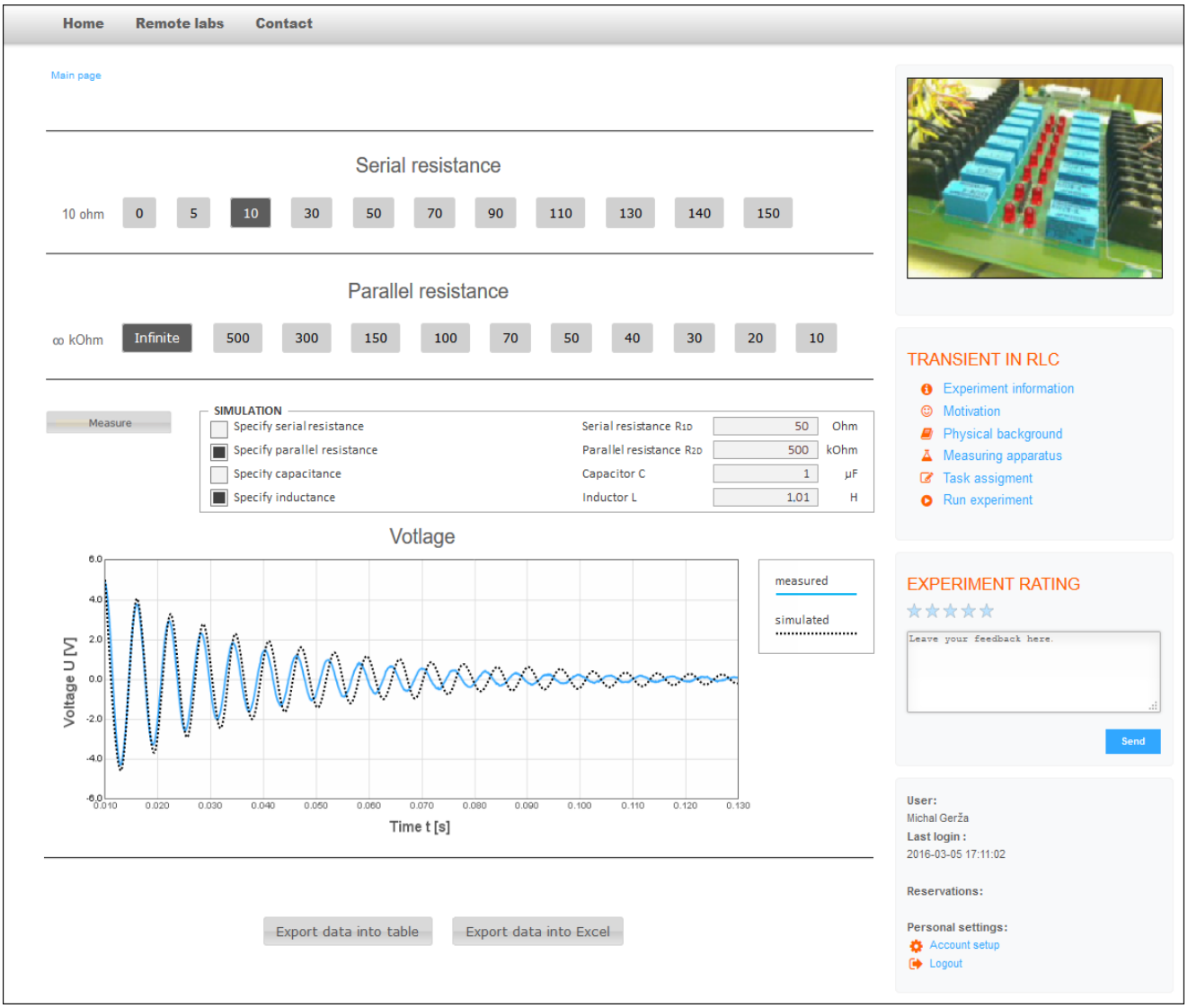


Figure 9. Web page of the ISES remote experiment "Transients in RLC" running together with the simulation process represented by the dotted black curve interlaced with a real measurement of the voltage in the parallel circuit.

and accept the 5th order approximation as the solution. This is the default solver for ODEs because it provides excellent performance in most situations. The solver interpolates using two steps of the bootstrapping algorithm applied to the Hermite interpolation (this gives order 5 interpolation). The interpolation scheme is very convenient because it significantly optimizes the number of evaluations of the ODE rate. The computational load is largely determined only by the tolerances and not by the reading step (Gonze, 2013). These solvers generate data, which are passed to the PSM output interface.

3.5 Concurrent data visualization

When the simulated data are received from the EJS solver to the PSM, the MS unit performs the synchronization with the concurrent real RE. Finally, the measured and simulated data are sent to the client's web page to the chart. Differences should be obvious, that is when we compare interlaced curves of the real and calculated voltage in the circuit. The web page visualizing both voltage representations is shown in Figure 9. This example simulation process allows setting of the damping resistors R_{1D} and R_{2D} both separately and by the buttons too for serial and parallel resistances. The optional values of the capacitance and inductance can be entered to modify the process. The real curve indicates R_{1D} = 10Ω, R_{2D} = Infinity and L = 1H, whereas the simulated curve indicates R_{1D} = 50Ω, R_{2D} = 500KΩ and L = 1.01H to observe obvious differences between the physical experiment and its defined mathematical model.

4 Conclusions

The paper introduced a new module designed for the phenomena simulation. Its advantage is perceivable in a creation of the simulation running concurrently with a real experiment in the ISES remote laboratory.

This module was integrated into the Measureserver unit. It allows clients its activation when the simulation process is demanded to complement the ISES remote experiment for the purpose of providing an alternative to observe differences between the mathematical and physical model of a studied phenomenon.

The simulation module should notably help students to improve their learning procedure concerning a better understanding of a given taught subject.

We formulated the following conclusions.

- The experimenting provided by the ISES remote laboratory is a new method of teaching and learning in comparison with traditional forms.
- The Measureserver unit is a core part used for measurement, data processing and communicating among clients and the ISES physical modules.
- The phenomena simulations module is a feature providing means to realize simulation process running together with the real experiment.

Acknowledgement

This work was supported by the European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2. 1.00/03.0089 and also by the Internal Grant Agency of Tomas Bata University in Zlín under the project No. IGA/FAI/2016/017.

References

A. S. Drigas, J. Vrettaros, L. G. Koukianakis, and J. G. Glentzes. 2006. *A Virtual Lab and E-Learning System for Renewable Energy Sources*. WSEAS Transactions on Computers 5, pp. 337-41. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-33645137921&partnerID=tZOtx3y1

Francisco Esquembre. *Easy Java/Javascript Simulations* [online]. Universidad de Murcia, Murcia, Spain: Ejs Wiki, 2015 [cit. 2016-03-17]. Available: http://www.um.es/fem/EjsWiki/Main/HomePage

Didier Gonze. *Numerical methods for Ordinary Differential Equations*. In: Teaching activities [online]. Belgium: Université Libre de Bruxelles, 2013 [cit. 2016-06-11]. Available: http://homepages.ulb.ac.be/~dgonze/TEACHING/numerics.pdf

R. Hamid and S. A. Modammed. 2010. *Remote Access Laboratory System for Material Technology Laboratory Work*. In International Conference on Engineering Education, pp. 311-16. Available: http://www.scopus.com/inward/record.url?eid=2-s2.079958730131&partnerID=tZOtx3y1

M. Krbeček, F. Schauer, and F. Lustig, *Easy Remote ISES - Development Environment Remote Experiments*, Innovations 2013. USA, 2013, pages 81-100. Available: http://www.ises.info/ oldsite/ clanky_pdf/Easy_Remote_ISES_2013.pdf

M. Krbeček, F. Schauer, and K. Vlček. *Communication Requirements of Laboratory Management System*. In: Latest Trends on Systems: Proceedings of the 18th International Conference on Systems, 2014, pp. 686-691. ISBN 978-1-61804-244-6.

F. Lustig. *Internet School Experimental System ISES* [online]. Prague, Czech Republic, 2009 [cit. 2014-06-05]. Available: http://www.ises.info/index.php/en/systemises

P. Zeman. *Software environment for control of remote experiments*. Ostrava: VŠB-Technical University of Ostrava, Czech Republic, 2011.

P. Zeman. *Software environment for integration of measured data from remote laboratory and simulation*. Ostrava. VŠB-Technical University of Ostrava, Czech Republic, 2012.

Development of a Hardware in the Loop Setup with High Fidelity Vehicle Model for Multi Attribute Analysis

Jae Sung Bang[1] Tae Soo Kim[1] Suk Hwan Choi[1] Raphael Rhote-Vaney[2] Harikrishnan Rajendran Pillai[2]

[1]Eco-Vehicle Control System Development Team, Hyundai Motor Group, South Korea, aeromec@hyundai.com
[2]MBSE Engineering Services, Siemens PLM, USA, raphael.rhote-vaney@siemens.com

Abstract

This paper describes a novel model based real time simulation approach to test, validate and calibrate electronic controllers for Hybrid Electric Vehicle (HEV) applications. The performance of the Hybrid Control Unit (HCU) needs to be evaluated on multiple vehicle attributes such as fuel economy, acceleration and drivability objectives. The multi attribute evaluation requires a higher level of detail for the vehicle simulation model where the energy flow and drivetrain dynamics are represented accurately. Given the high mechatronic content and the strong interactions among the various controllers in HEVs, it becomes necessary to simulate many of the vehicle controllers on the real time platform. The higher fidelity vehicle model coupled with the realistic behavior model of the controller network poses challenges in setting up the real-time Hardware In the Loop (HIL) test platform where the vehicle level attributes can be studied. The real time simulation setup process, its challenges and the methods used to overcome these challenges are described in this paper.

Keywords: hardware in the loop, Amesim, hybrid electric vehicle

1 Introduction

The expectations from the consumer has transformed radically in the recent years from the advent of enhanced driver support, better fuel efficiency and improved powertrain technologies. Automotive manufacturers and suppliers are confronted with ever greater complexity as a result of increasing numbers of products and options, shorter technology cycles and the increasing pressure to innovate. At the same time they need to balance the needs and demands of customers, investors, regulators, non-governmental organizations and even the general public (Pwc, 2014). The passenger vehicles are being transformed to mechatronic machines with high electronic and software contents. Companies cannot afford to test such complexity in a hardware prototype thoroughly because of the extremely high costs associated with design changes far down the development cycle (Boehm,2005) and the extraordinary lead time associated with such a task.

Companies are using model-based simulation approaches to design and test such high complexity mechatronic systems. Testing the controller unit comprehensively before testing in the prototype requires real-time Hardware in the Loop (HIL) test platforms. Most of the existing literature in the HIL area (Nabi, 2004; Ramaswamy, 2004; King, 2004; Allende, 2015; Hafiz, 2015; Bovee, 2015; Isermann, 1999; Basrah, 2015; Wei, 2004) refers to the usage of HIL testing in the context of controller logic validation and testing.

It is difficult to find a study that focuses on real time HIL simulation of a complete closed loop vehicle model that captures both fuel economy and drivability phenomenon accurately. It is in this scenario, that the process we have created (i.e. performing vehicle level multi-attribute analysis with controller in the loop on HIL) becomes unique.

The paper describes the process of real time closed loop HIL validation of a hybrid electric vehicle model with the objective of validating the supervisory hybrid control algorithm on the basis of both fuel economy and drivability characteristics of the vehicle.

The second section in this paper describes the hybrid electric vehicle architecture, scope of the multi-attribute analysis and the load cases used to study the multiple attributes. The Amesim© model used to describe the Hybrid Electric Vehicle (HEV) powertrain, the controller model architecture and the process followed for modeling and validating the system is described in the third section. The real-time model generation process and the partitioning of the model to execute on the multi-core platform to optimize execution performance are discussed after that. This is followed by results discussion and conclusions.

2 Multi-attribute analysis of hybrid electric vehicles

This section describes the architecture of the hybrid electric vehicle under consideration, functionality of the Hybrid Control Unit (HCU), other controllers and the

test cases. Within the scope and context of this paper, multi-attribute analysis refers to the simultaneous analysis of fuel efficiency and drivability characteristics of the vehicle. The entire simulation setup has been created in order to study the impact of the HCU strategy on the drivability and fuel efficiency related aspects of the vehicle.

The configuration of the vehicle in this paper is referred to as a Transmission Mounted Electric Device Hybrid Electric Vehicle (TMED HEV). As seen in Figure 1, the architecture is composed of a Hybrid Starter and Generator (HSG), an engine, an engine clutch, an electric motor and a transmission. The HSG is connected to the engine by a pulley and the main role of the HSG is to start the engine, control the engine speed to engage the engine clutch and charge the battery by using engine power. The engine clutch plays an important role in the mode of the vehicle such that when the clutch is disengaged, the engine power is not transferred to the driving wheels and only motor power drives the vehicle. This setting is

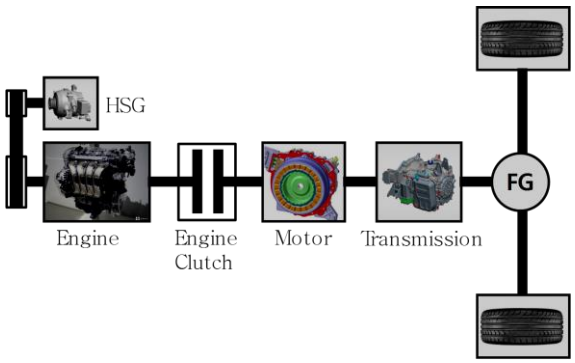


Figure 1. HEV Architecture

known as the Electric Vehicle (EV) mode. When the clutch is engaged, the engine power and the motor power drive the vehicle in HEV mode. During the HEV mode, the assistance by the electric motor ensures that the operating point of the engine is maintained to have optimum fuel efficiency. The 6-speed automatic transmission is similar to a conventional transmission except for the fact that it does not have a torque converter to reduce energy loss.

The optimal control of the operation of the vehicle is managed by an array of control units. Out of these, the HCU is the supervisory controller that computes the torque demanded by the driver and optimally distributes it among the engine, electric motor and the HSG. The torque demand from the driver is estimated by the HCU by considering multiple factors such as the accelerator pedal position, the brake pedal position, vehicle speed, gear ratio and torque interventions by other controllers. For example, if the battery state of charge is sufficiently high and the torque demand is less than the maximum motor torque at the current motor speed, then the HCU generally opts for the EV mode of operation in which all of the demanded torque is supplied by the electric motor. On the other hand, in the case of HEV mode, the HCU distributes the demanded torque between the electric motor and the engine by considering the brake specific fuel consumption of the engine and the efficiency of the electric motor at that particular operating point. Hence it can be seen that the HCU plays a critical role by ensuring the driver demands are met and simultaneously ensuring optimal fuel efficiency.

In addition to the supervisory HCU, there are other subsystem level controllers (explained in the next sections) that also have to be modeled to an appropriate level of detail to ensure the simulation results are realistic and comparable to vehicle test data. For instance, the Engine Management System (EMS) controls fuel, air and spark in order to produce the commanded torque from HCU. If this feature is not modeled correctly, it can lead to significant deviations in the prediction of fuel economy and battery state of charge.

In order to validate the performance of the model across the two attributes of drivability and fuel economy, the following test cases are utilized:

- For fuel economy analysis:
    - FTP drive cycle
    - US06 drive cycle
    - NEDC drive cycle
- For drivability analysis:
    - Fixed gear HEV mode tip & creep
    - HEV/EV mode change

3 System modeling for multi-attribute analysis

This section highlights the modeling aspects of the plant model using Amesim© software and the controller models using MATLAB®/SIMULINK®/STATEFLOW®. A brief description of the process followed to validate the models is also provided.

Since the HCU is to be tested against Fuel Economy and Drivability requirements, a dynamic model of the vehicle using Amesim© is developed with the physics needed to capture the energy flows and conversion from fuel to mechanical and electrical energy. In order to also address drivability requirements, the level of details for the description of the elements involved is chosen to capture natural frequencies in the 0 to 20Hz range which corresponds to the frequencies that can be felt by the driver as seen in Figure 2

Initially, the vehicle model is developed focusing on these considerations, the real-time capabilities not being part of the requirements considered at the time. In order to capture the targeted frequency content, all the mechanical elements, whose modes are known to fall within that range, are included in the model (i.e. drivetrain, engine 3D block and mounts, suspension and

chassis). The 6DOF engine block on its mounting system and the corresponding rigid body modes are shown in Figure 3.

In order to use the vehicle model for predicting Fuel Economy, the efficiencies and energy losses of the main components should be modeled. Being able to track and account for the power flows is key to understanding how the fuel energy is converted to the mechanical energy delivered to the wheels. Once this torque/energy balance is achieved, the engine brake torque and speed are used to calculate the corresponding fuel consumption. The BSFC map used to predict the engine fuel consumption

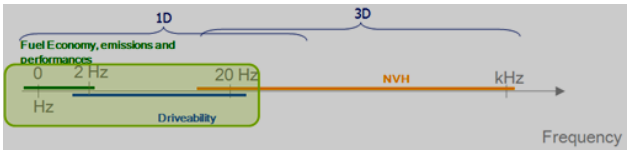


Figure 2. Frequency range of model

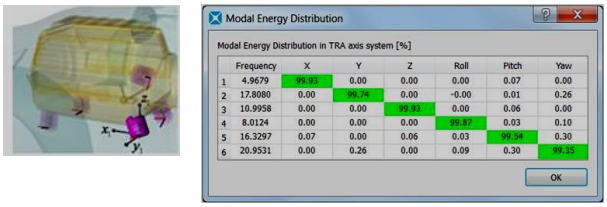


Figure 3. Modal energy distribution

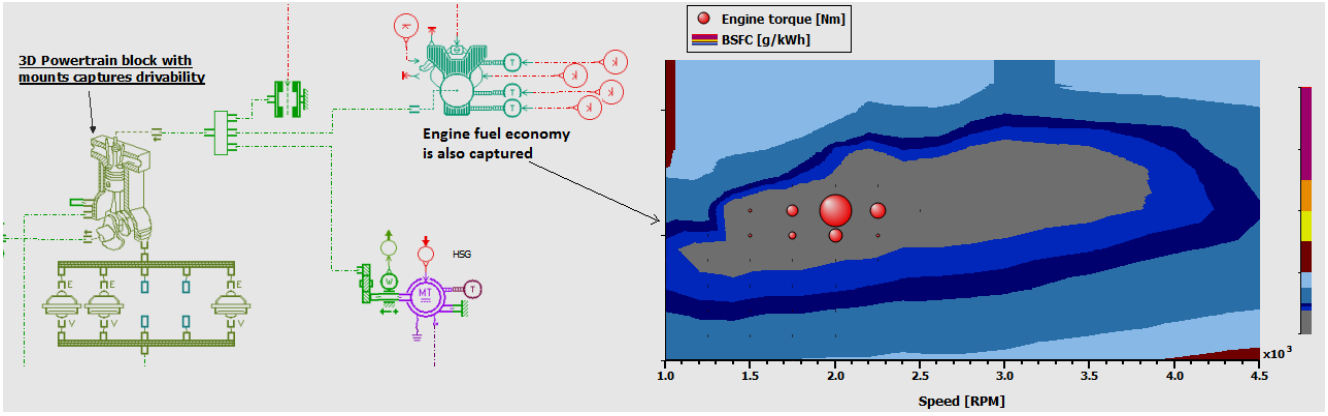


Figure 4. Model captures drivability and fuel economy

from engine torque and speed is shown in Figure 4.

It is noteworthy that the plant model architecture mirrors that of the physical hardware so as to capture the functions and physics needed to address the requirements of interest and capture the corresponding physical phenomena. The complete model being sizeable, it is not presented in this paper but it shall be introduced during the conference.

Now moving on the controller modelling aspects, a summarized version of the overall vehicle controller network architecture is outlined in Figure 5. The figure corresponds to the system that is modeled within the scope of this particular project. It can be seen that the HCU is linked to the rest of the controllers through Controller Area Network (CAN) and these controllers in turn are connected to each other and to the plant hardware.

As stated previously, HCU is the high level supervisory control unit that manages the torque split between the internal combustion engine and the electric motor according to the operating conditions. The HCU computes and sends the main signal of each controller such as the engine torque command, the motor torque command, the battery charge and so on. The other controllers carry out the command from the HCU. The Engine Management System (EMS) controls the fuel quantity, air quantity and ignition timing in order to realize the torque command from the HCU. The EMS also estimates the amount of fuel consumption. The Motor Control Unit (MCU) controls the electric motor by controlling the current. The MCU also ensures a reduction in driveline oscillations by appropriately controlling the motor torque and also shifts the operation point of the engine to achieve better fuel efficiency. The HSG, which controls the engine speed for engine clutch engagement and which is used to charge the battery, is also controlled by the MCU. The Transmission Control Unit (TCU) determines the shift point and provides the commands to actuate the clutches and the brakes within the automatic transmission. The Battery Management System (BMS) monitors the state of the battery at all times and decides the charging/discharging limits based on the operating point. The BMS also computes an estimate of the state of charge (SOC) of the battery. The Anti-lock Braking System (ABS), within the context of this project, computes the total braking torque required based on the driver input and then splits them appropriately so that a certain percentage of the total braking requirement is supplied using the hydraulic brakes and the rest is produced by the motor as regenerative torque. The above controller network structure (excluding the HCU), along

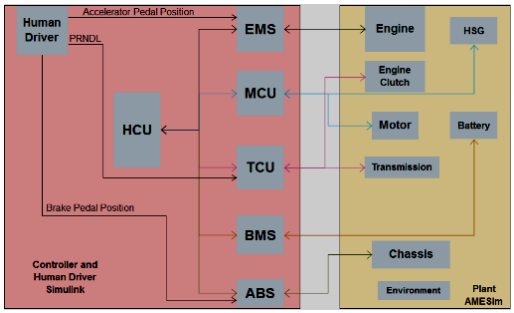


Figure 5. Model Architecture

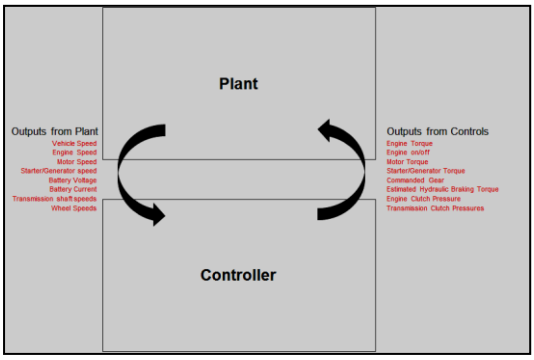


Figure 6. Plant-Controls Interface

Amesim© plant modeling is carried out at the component level first, which is then assembled to obtain the complete vehicle model. Validation is also ensured at both the component and system level of the plant. The controller models are also built and validated in a systematic way, from unit level to system level The process of controller model building – from unit level logic, to complete 'X'CU, to the entire controller is highlighted in Figure 7.

The validation process also follows the development process – open-loop unit level validation is followed by open-loop system level validation. Once this is completed for both the plant and control model, closed-loop validation is performed on the desktop. The final validation step is performed on the HIL bench using the HCU hardware. The entire validation process is concisely represented in Figure 8.

4 Real-Time Model Generation and Setup

Initially developed to capture drivability phenomena like clutch Judder or Shift surge, the Amesim© model contains excessively high natural frequencies and very small time-constants that cannot be handled by a fixed

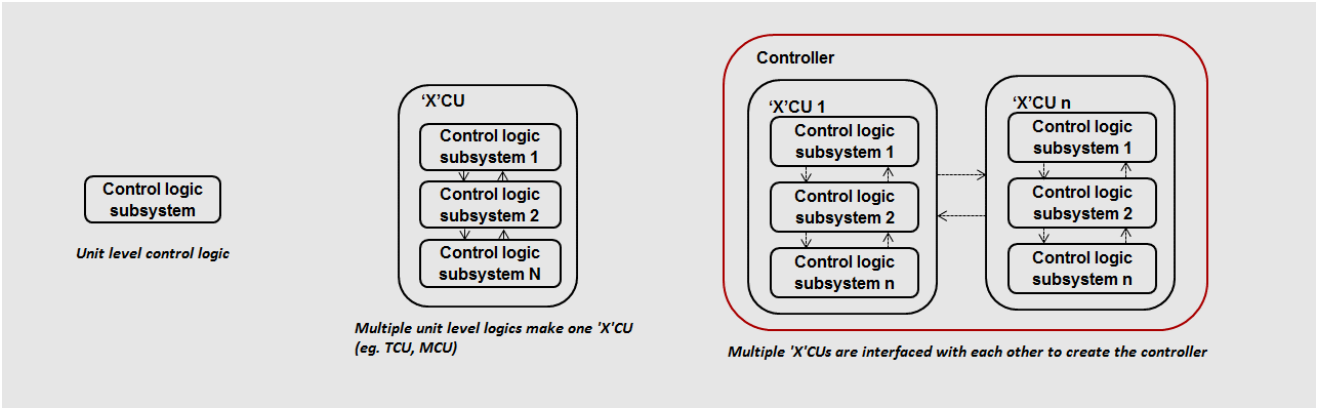


Figure 7. Controller modeling approach

with a simplified implementation of the algorithm which includes the main function of each controller, is modeled in Simulink.

Since the vehicle plant model is developed using the Amesim© software and controller models are developed using the MATLAB®/SIMULINK®/STATEFLOW® suite, it is very important to define the interface between these two entities early on in the development process. Effort is taken to ensure that the virtual interface is similar to the actual interface implemented on the real vehicle, to the extent possible. The signal interface between Amesim© (plant model) and MATLAB® (controller model) is shown in Figure 6.

Modeling and validation of the plant model and controls are done following a systematic procedure with confirmation and validation ensured at each stage. The time-step solver. In order to use in an HIL setup several steps have to be taken for the model to get rid of the unnecessary high frequency while keeping the lower frequency content used to capture drivability phenomena. The first step is to identify the largest frequency or mode and the main contributing states. Fig. 9 shows the Modal Projection Tool that is used to that effect.

In a second step the part of the model contributing to this mode is simplified. Then the new results are compared to that of the original model to validate that differences in time-domain and frequency-domain are acceptable while the model is being significantly reduced. In the present case, the inertia of the HSG, connected to the engine flywheel inertia via a very stiff and over-damped shaft, generates the 47MHz mode shown in Figure 9. This may not be a problem for a

variable time step solver but that will be a problem for a fairly typical fixed time-step. In this example the two inertias can be lumped together and the very high mode disappears. This process is then repeated until all the unnecessary high frequencies have been removed. Figure 10 shows natural frequencies of the system after model reduction and the drivability results after vehicle model reduction is plotted in Figure 11. The frequency, magnitude, vehicle speed, and acceleration are very similar to vehicle test data, and the results are sufficiently accurate to predict the behavior of the powertrain.

The following subsection highlights certain aspects on the controller side that needs to be taken care of before deploying on a real-time bench.

The different controllers are to be executed at various sample periods and also the communication intervals among the different controllers vary. All of these factors have been taken into consideration for controller modeling. For the purpose of desktop Model in Loop (MIL) simulation, the controller models and the plant model (Amesim© S-function) are within the same Simulink .mdl file. However, for the HIL simulation, the plant model (Amesim© S-function) is simulated in a separate mdl file. The structure of HIL model is shown in Figure 12.

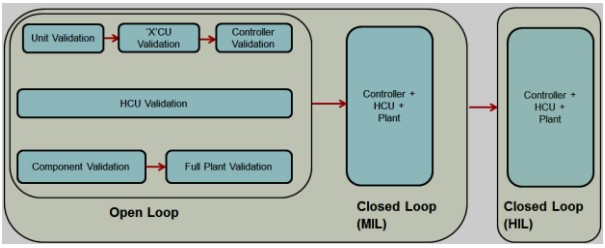


Figure 8. Validation Approach

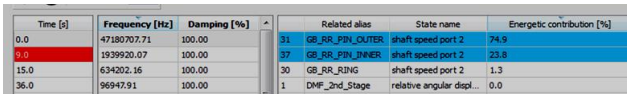


Figure 9. Natural frequencies before Transmission model reduction

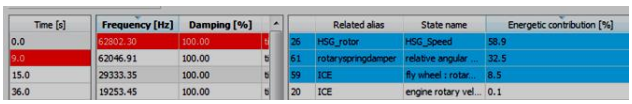


Figure 10. Natural frequencies after Transmission model reduction

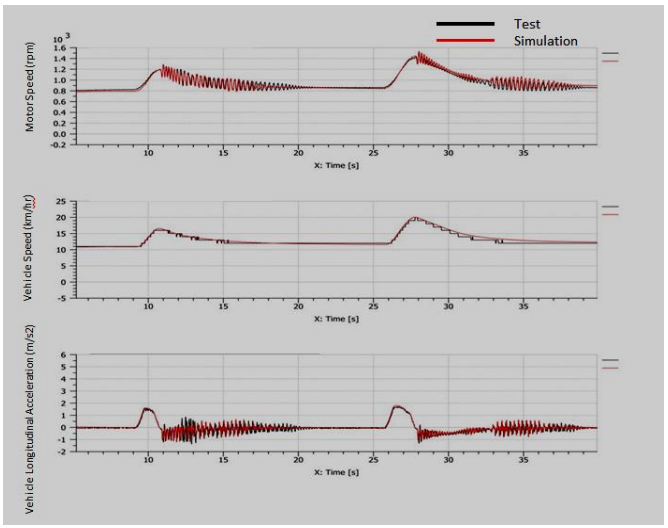


Figure 11. Drivability Results

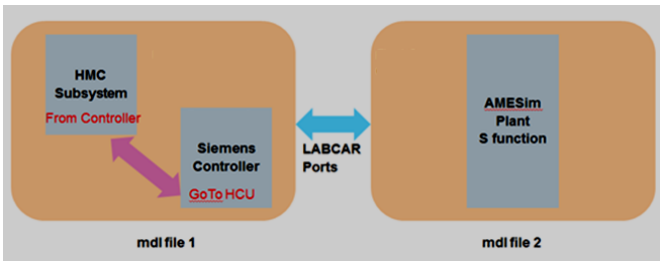


Figure 12. HIL model structure

Here, mdl file 1 contains the controller blocks. This model is executed using a fixed time step solver. Internal clocks within this model take care of the separate trigger rates required for the different control systems. The second mdl file contains the plant model S function and this file is executed at a fixed time step which is smaller than the controller models. Since the system model is split into two separate mdl files for

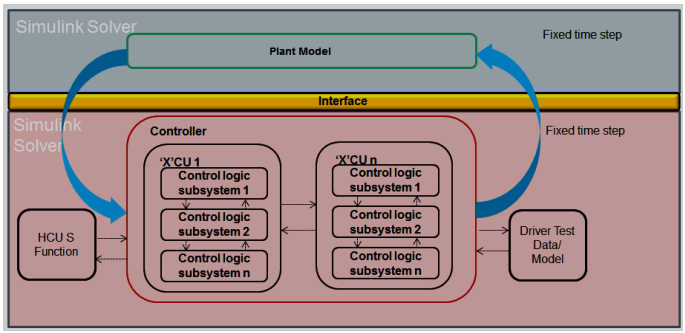


Figure 13. Closed loop MIL model structure

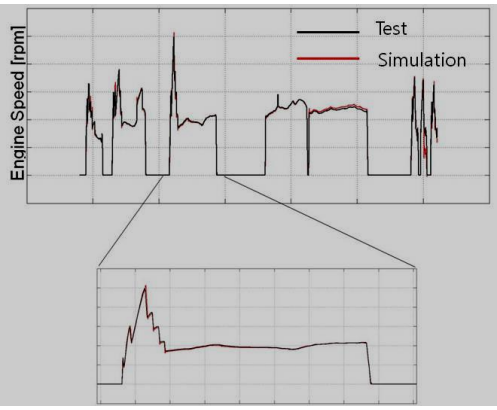


Figure 14. MIL Results

controller models and plant model, it is possible to simulate them on two separate cores on the HIL bench.

Once the real time plant model and controller models are ready, the closed loop setup is validated against test data on desktop. For this purpose, an S function of the HCU is utilized.

The overall model structure used to test the system closed loop on the desktop is outlined in Figure 13. It can be seen that the entire controller model components are in the Simulink environment and the plant model is implemented as a MATLAB® S function. The entire system is simulated in fixed time step using the Simulink solver. The HCU is also implemented as a MATLAB® S function. This setup enables a closed loop fixed step simulation on the desktop before it is deployed on the HIL bench.

Figure 14 shows the result of the MIL simulation using the HCU s function. The engine speed error between test and simulation is extremely small. This close behavior between model and test data on desktop ensures the accuracy of the model before deployment on HIL.

5 HIL Test Bench Results

The novel simulation model on the HIL bench is used to analyze the impact of HCU algorithm change or calibration change on the fuel efficiency and drivability aspects of the HEV. The HIL bench enables evaluation of the performance of the HEV without real vehicles thereby reducing the development time significantly.

The HIL bench used for the multi-attribute analysis and the verification/validation of HCU is shown in Figure 15. The HIL bench is composed of the real-time PC, the operating PC, I/O board and HCU. Using LABCAR© software, the vehicle model and controller model suggested in this paper are built and the INCA© software is used for measurement of signals and calibration.

Some of the issues related to HIL deployment is outlined in this paragraph. The main issue that can cause problems during HIL execution is model overrun. This means that all of the computations associated with the system model are not being completed within the allocated fixed time step. When this is encountered, the model can be split and run on multiple cores. This would entail having two separate mdl files with LABCAR© ports for communication between them. Simulation debugging can also be a challenge in this case because of the number of constituents in the system – plant, controller models, HCU hardware, interfaces, different cores etc.

The results presented in this section reflect the Sonata HEV vehicle. Various versions of the prototype HCU are used to correlate the test result and the simulation result. In the figures, red line corresponds to the test data from the real vehicle and the blue line corresponds to the simulation data.

Figures 16, 17 and 18 show the full closed loop simulation results obtained from HIL simulation with HCU hardware connected. The HCU, vehicle and the controller models have the same calibration values. A subsection of the result from FTP drive cycle is shown in Figure 16. For reasons of security, the result for the entire time range is not provided. The vehicle speed error between the test data and simulation is below 2 km/hr. Since the error is less than the acceptable maximum speed error value of 3.2 km/hr, other variables of the simulation is expected to match the performance of the Sonata HEV for the FTP drive cycle. The engine torque error between the test and simulation is below 3%. The difference in the timing of engine on/off point is less than one second which is very small considering the whole range of the FTP drive cycle (1300s). The small difference in the engine on/off signal arises from the error in vehicle speed, SOC, and so on. The errors for SOC and the fuel consumption are also below 3%. Since all errors are within acceptable limits, the prediction of the fuel economy is possible for the FTP drive cycle. Figure 17 shows a subsection of the results of the Highway drive cycle. The errors for vehicle speed, engine torque, battery SOC and the fuel economy is below 3%. It is to be noted that the simulation model used for validating the highway drive cycle is exactly the same as the one used for FTP. The characteristics of the FTP and the highway drive cycles are vastly different. The FTP drive cycle can be considered to be somewhat mild in terms of the rates of acceleration and braking. The highway drive cycle, on the other hand is aggressive. Since the simulation model is able to match the vehicle performance for both of these cycles, any other cycle's fuel economy can be predicted. Finally, the result of NEDC drive cycle is shown in Figure 18. The results are sufficiently similar to the test data of the real vehicle and therefore the fuel economy for the NEDC is predicted by the developed model.

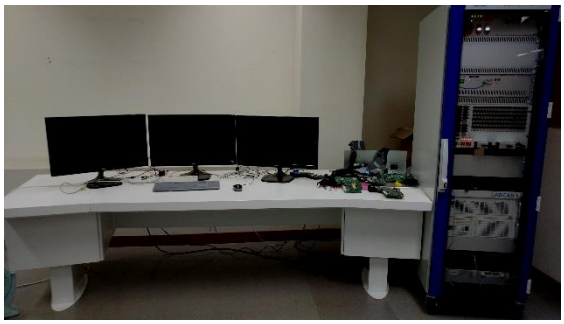


Figure 15. HIL bench for HCU

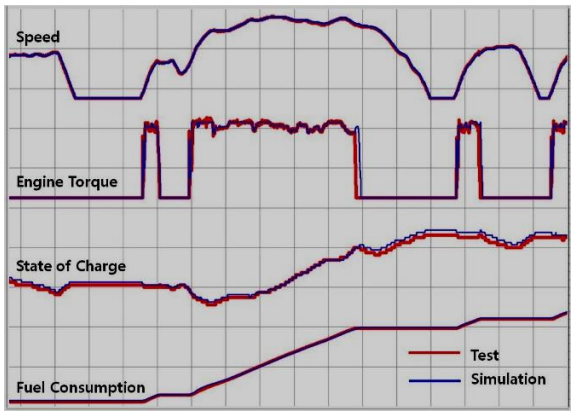


Figure 16. HIL FTP cycle results

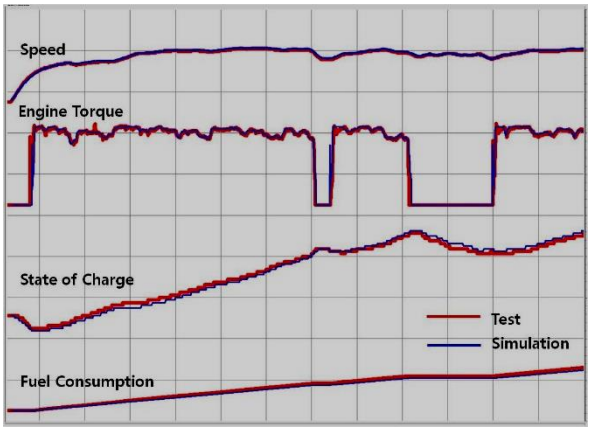


Figure 17. HIL Highway cycle results

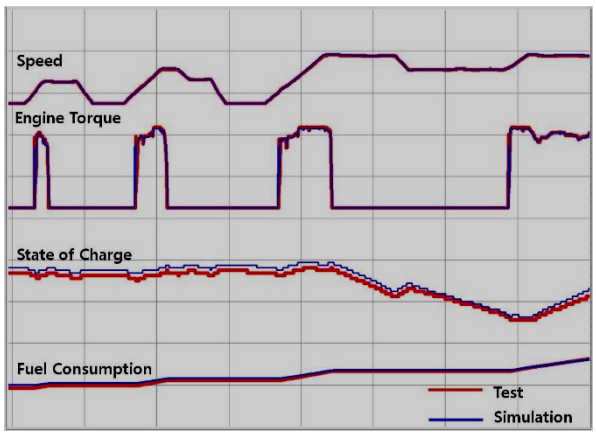


Figure 18. HIL NEDC cycle results

6 Conclusions

The results of the HIL simulation show a very close correlation between the real vehicle and the model. This confirms that closed loop vehicle simulation on an HIL bench can be used to validate multiple attributes like drivability and fuel economy. This process opens up new opportunities for similar multi-attribute studies on the HIL bench. This approach is novel when compared to existing studies that tend to focus on a single attribute using lower fidelity plant models for controls logic validation.

This setup enables Hyundai Motor Company (HMC) to quickly test the impact of different HCU algorithms and calibration values on the vehicle drivability and fuel economy. The overall development time for the HEV has been reduced and the performance of the HEV has been improved.

The vehicle model presented in this paper does not include thermal effects on the powertrain and also has limited predictive capabilities of the accessory loads.

As a next step, HMC and Siemens would like to include the effects of the temperature on the engine, electric motor and the battery. An improved model which includes the electrical energy consumption by the air conditioner, electric oil pump and the electric power steering will provide a better estimate of the energy consumption. In addition, if the road geometry and the vehicle lateral dynamics are added to the model, a variety of other test scenarios can be simulated accurately.

Acknowledgements

We would like to thank other team members (HMC, Siemens PLM colleagues in Lyon, Detroit, Chennai and Seoul) who indirectly contributed to this paper.

References

M. Allende, P. Prieto, B. Hériz, J. M. Cubert, and T. Gassman. Advanced Shifting Control of a Two Speed Gearbox for an Electric Vehicle. In *The 28th International Electric Vehicle Symposium and Exhibition*, pages 118-128, Korea, 2015.

M. S. Basrah, E. Velenis, and D. Cao. Four wheel torque blending for slip control in a hybrid electric vehicle with a single electric machine. In *ICSEEA*, pages 19-24, IEEE, 2015.

B. Boehm and V. R. Basili. Software defect reduction top 10 list. *Foundations of empirical software engineering: the legacy of Victor R. Basili*, 426(37):135-137, 2005.

K. Bovee, A. Hyde, M. Yatsko, M. Yard, M. Organiscak, B. Hegde, J. Ward, A. Garcia, S. Midlam-Mohler, and G. Rizzoni. *Plant Modeling and Software Verification for a Plug-in Hybrid Electric Vehicle in the EcoCAR 2 Competition*. SAE Technical Paper, 2015.

F. Hafiz, P. Fajri, and I. Husain. Effect of brake power distribution on dynamic programming technique in plug-in series hybrid electric vehicle control strategy. In *IEEE*

Energy Conversion Congress and Exposition (ECCE), pages 100-105, IEEE, 2015.

R. Isermann, J. Schaffnit, and S. Sinsel. Hardware-in-the-loop simulation for the design and testing of engine-control systems. *Control Engineering Practice*, 7(5):643-653, 1999.

P. J. King, and D. G. Copp. Hardware in the loop for automotive vehicle control systems development. In *UKACC Control 2004 Mini Symposia*, pages 75-78, IET, 2004.

R. Mura, V. Utkin, and S. Onori. Energy management design in hybrid electric vehicles: A novel optimality and stability framework. *IEEE Transactions on Control Systems Technology*, 23(4):1307-1322, 2015.

S. Nabi, M. Balike, J. Allen, and K. Rzemien. *An overview of hardware-in-the-loop testing systems at Visteon*. SAE Technical paper, 2004.

D. Ramaswamy, R. McGee, S. Sivashankar, A. Deshpande, J. Allen, K. Rzemien, and W. Stuart. *A case study in hardware-in-the-loop testing: Development of an ECU for a hybrid electric vehicle*. SAE Technical Paper, 2004.

X. Wei. *Modeling and control of a hybrid electric drivetrain for optimum fuel economy, performance and drivability*. Doctoral dissertation, The Ohio State University, 2004.

R. Hanna, and F. Kuhnert. *How to be No. 1: facing future challenges in the automotive industry*. PwC Autofacts, 2014.

From Low-Cost High-Speed Channel Design, Simulation, to Rapid Time-to-Market

Nansen Chen[1] Mizar Chang[2]

[1]SV Div., Home Technology Development, MediaTek Inc., Taiwan, nansen.chen@mediatek.com
[2]CTE Div. II, Analog Design and Circuit Technology, MediaTek Inc., Taiwan, mizar.chang@mediatek.com

Abstract

Leadframe packages are always adopted as the low-end devices. When the low-cost channel including the leadframe package and the two-layer PCB is required for high-speed digital signaling over 1 Gb/s, the iteration of full channel simulation and analysis with reliable EDA tools should be taken before the device is rolled out. Different channel designs were characterized in the frequency domain using the 3-D full-wave electromagnetic field solver to analyze the bottleneck of channel performance. Comparison of the full channel S-parameters, the channel with the proposed DDR3 memory controller package suffers less insertion loss. The chip-package-board co-simulations in the time-domain using the chip HSPICE netlists and full channel S-parameters for the DDR3 data accessing at 1.2, 1.4, and 1.6 Gb/s were taken and demonstrated that the channel including the proposed package design had larger timing and voltage margins, and less jitter, overshoot and undershoot, which all conform to JEDEC Standard. The waveform measurement also verified the same prediction that the DDR3 memory controller encapsulated in the modified E-pad LQFP package achieved no cost impact and enough timing margin up to 1458 Mb/s. The performance of mature leadframe packages can be promoted if the careful package designs are taken.

Keywords: DDR3, E-pad, LQFP, return path, S-parameters, jitter, eye diagram, JEDEC

1 Introduction

Before 2009, the year of DRAM transition from DDR2 to DDR3, several famous fabless semiconductor companies predicted that the DDR3 memory controller should be designed and encapsulated with the flip-chip ball grid array (BGA) package because the wire-bonding packages induce large inductance or impedance that is harmful to the single-ended DDR3 signals accessing over 1 Gb/s. The low price is always the king for the consumer electronics market, such as LCD TVs, BD players and broadband Wi-Fi routers, even though the low-cost 2-layer PCB would be implemented. In the following years, many design guides and studies have been proposed to recommend wire-bonding or flip-chip BGA packages for the DDR3 memory controller (Micron Tech., 2009; Texas Inst., 2014; Shah, 2012; Synopsys Inc., 2009). However, few papers presented the investigation of DDR3 memory controller encapsulated with the wire-bonding leadframe packages. In this paper, several passive channels were designed and studied whether the memory controller with the exposed die-pad (E-pad) low-profile quad flat package (LQFP) was acceptable to access data rate up to 1.6 Gb/s using the chip-package-board co-simulation in frequency and time domains. The effects of different return paths in the memory controller package were characterized with the 3-D full-wave electromagnetic field solver to demonstrate the bottleneck of channel performance. Finally, the DDR3 waveform measurement in the practical platform was taken matching the previous co-simulation prediction of signal integrity. The reliable simulation tools are very important to analyze the channel performance for different designs that can predict the effects of nonideal return paths correctly and is helpful to finalize the channel design and expedite time-to-market.

2 Package Structures

Leadframe packages are made of the single-layer copper-based alloys. Thus, they are cheaper but suffer worse electrical performance, including larger cross-talk and energy loss due to longer parallel leads and bondwires, compared to the ball grid array (BGA) packages with at least two-layer substrate. Figure 1 shows the structure of exposed die-pad (E-pad) low-profile quad flat package (LQFP) with 256 pins. In order to increase the signal inputs/outputs interconnected between the chip and external devices on the PCB, all ground wires are bonded onto the ground bar, which connects with the E-pad through the bridges or the connecting bars. Then the E-pad soldered with the ground pad of PCB connects to the global ground. Another benefit for the E-pad is to promote the heat dissipation. If the ground wires are bonded onto the E-pad top surface, moisture may easily penetrate into the package through the interface between the molding compound and the E-pad. In addition, there is a reduction in the coupling strength of the E-pad to the

molding compound. As a result, an interface peel-off phenomenon caused by thermal stress may occur (Choi, 2002). Accordingly, the ground wires are disconnected from the E-pad. That is why all the ground wires from the chip shall be bonded onto the ground bar, which is elevated from the package bottom surface.

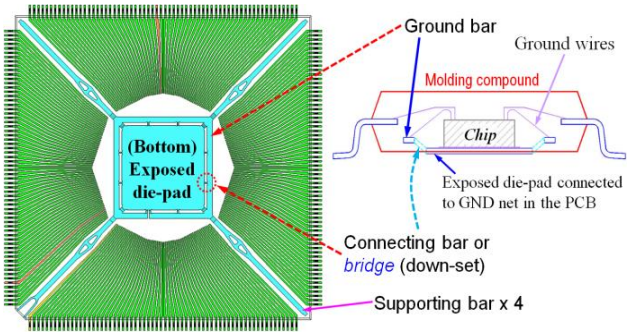


Figure 1. Top and side views of 256-pin E-pad LQFP package.

3 Co-Simulation Methodology

The passive channel design is critical to maintain the signal and power integrity of high-speed digital signals, especially for the package design. In order to ensure the acceptable channel performance, the iterative co-simulation using the reliable EDA tools were taken. Both the chip netlists of DDR3 memory controller and DRAM cascaded with the wideband channel S-parameters including the power and signal nets are modeled for the transient analysis, as shown in Figure 2. The chip input/output buffer information specification (IBIS) models are not recommended because those behavior models are less accurate for the signal speed over 1 Gb/s.

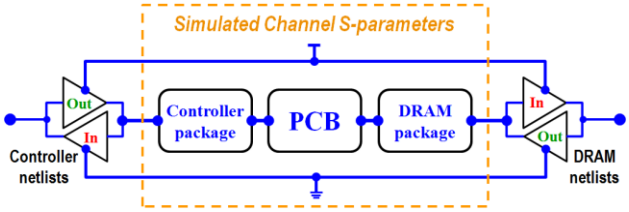


Figure 2. Channel models of DDR3 interface co-simulation.

4 Channel Analysis

The low-cost DDR3 channel configuration includes the memory controller package, the PCB and the DRAM package, as shown in Figure 3. Two facts the fabless chip companies are unable to change. The first fact is that the DRAM package type and ball pins are defined by JEDEC Standard (JEDEC Std., 2012). The second fact is that many original equipment manufacturers (OEM) always choose the 2-layer PCB rather than the 4-layer PCB due to 40% cost reduction, as listed in Table 1 (Chen, 2009). Finally, the fabless chip companies only can determine the memory controller package type. According to the package cost comparison listed in Table 2 (Chen, 2009), the adoption of leadframe packages can save up to 92−208% in package cost. The challenge is whether the full channel performance is acceptable for the data access over 1 Gb/s. In order to realize the performance difference of the memory controller encapsulated in the leadframe and the BGA packages, both full channel S-parameters including the DDR3 signal and the I/O power net (1.5 V) were extracted using ANSYS HFSS, a 3-D full-wave electromagnetic field solver, and then cascaded with the chip netlists for the transient analysis in Synopsys HSPICE. As demonstrated in Figure 4 obviously, the channel with the conventional E-pad LQFP has larger skew and insufficient timing margin compared to that with the BGA package. How to improve the E-pad LQFP performance became a must.

Table 1. Examples of PCB cost ratios in digital TV mother boards.

Type / Item	Model-X TV		Model-Y TV	
	2-layer PCB	**4-layer PCB**	**2-layer PCB**	**4-layer PCB**
	Cost Ratio	Cost Ratio	Cost Ratio	Cost Ratio
Controller chip	37.2%	34.7%	26.7%	24.9%
PCB	**9.3%**	**15.4%**	**9.1%**	**15.1%**
Other components	53.5%	49.9%	64.3%	60.0%

Table 2. Cost ratio comparison among different package types.

Package Type	Size (mm)	Cost Ratio	Remarks
E-pad LQFP216	26 x 26	**0.77**	216 pins.
E-pad LQFP256	30 x 30	**1.00**	Comparison baseline.
PBGA (2-layer)	27 x 27	**1.92**	With plating lines.
PBGA (2-layer)		**2.12**	Without plating lines.
PBGA (4-layer)		**2.12**	With plating lines.
PBGA (4-layer)		**2.31**	Without plating lines.
FC-BGA (4-layer)		**3.08**	Exclusive of bumping cost.

5 Improved Leadframe Package

Several simulations of full channel S-parameters were taken including the E-pad LQFP package with different ground bar widths, bridge widths and numbers. Figure 5 shows the partial pictures of E-pad leadframe packages with different bridge numbers. Finally, the bridge number is the key factor to improve the channel performance. As package modeled with different numbers of bridge shown in Figure 6, the simulation results indicated in Figure 7 that the package with more bridges suffers less insertion loss than that with fewer bridges. The improved amplitude is 3 dB at 1.3 GHz. The bridges connected between the ground bar bonded with the ground wires from the chip and the exposed pad (E-pad). The more the bridges, the smaller return loop

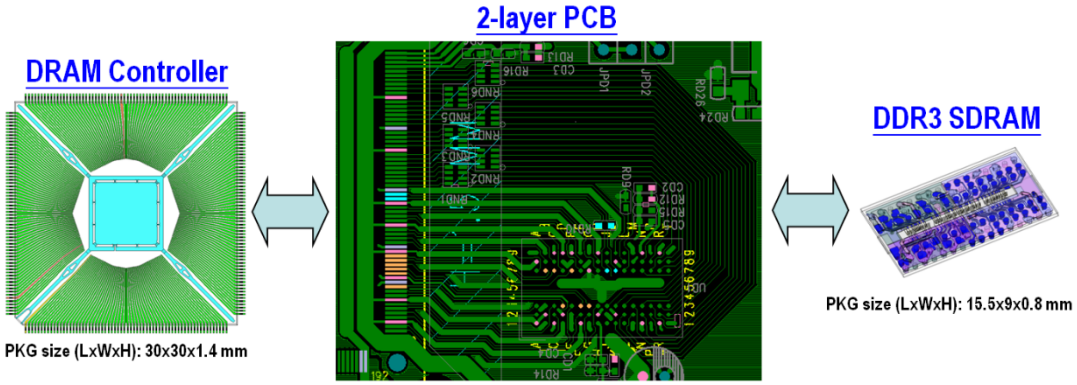
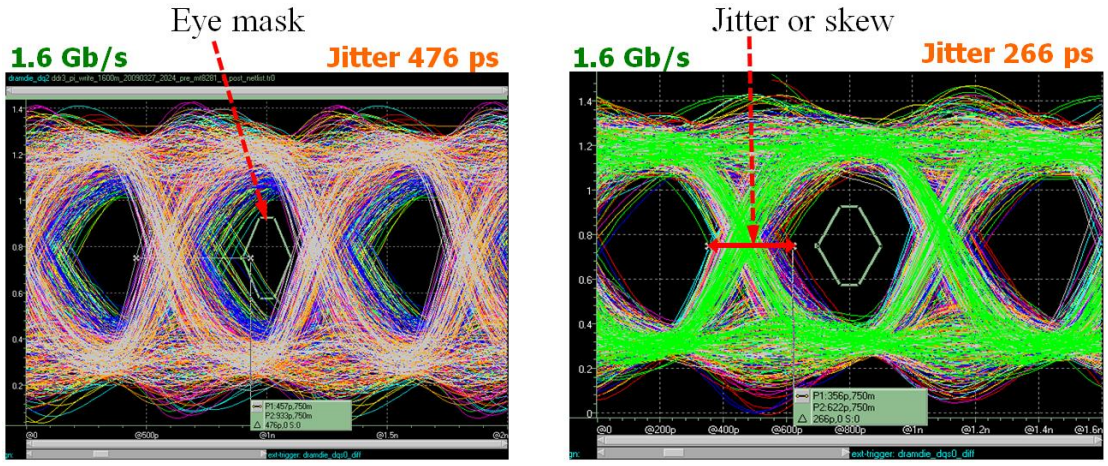


Figure 3. Low-cost DDR3 channel configuration.



(a) (b)

Figure 4. Simulated DDR3 eye-diagrams of overlapping 1-byte signals on DRAM chip side for writing data at 1.6 Gb/s. (a) Memory controller in conventional E-pad LQFP. (b) Memory controller in BGA package.

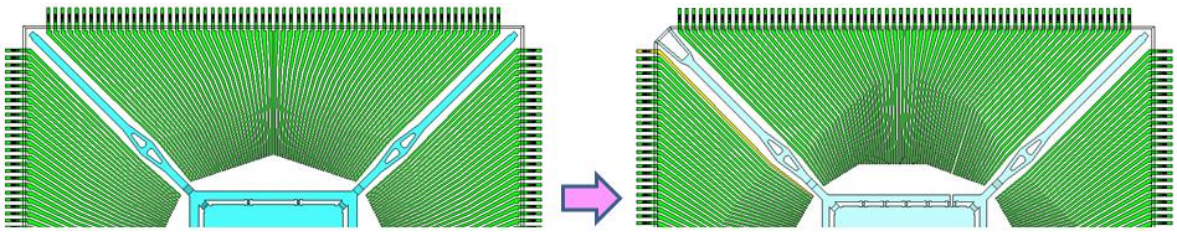
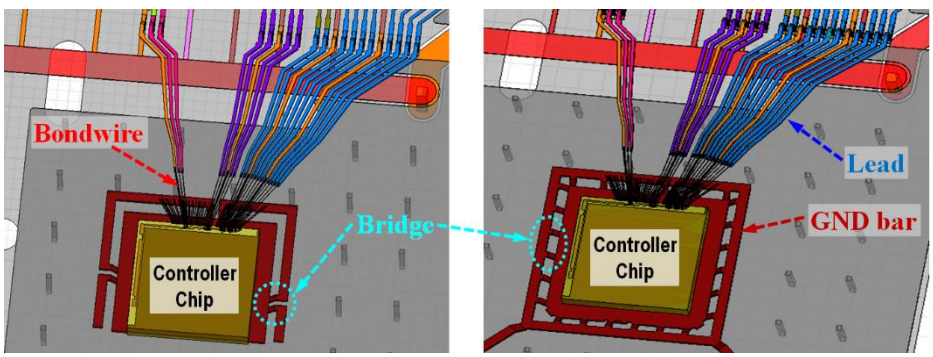


Figure 5. Modification of E-pad leadframe packages with different bridge numbers.



(a) (b)

Figure 6. Simulation models of memory controller packages. (a) E-pad LQFP with few bridges. (b) E-pad LQFP with many bridges.

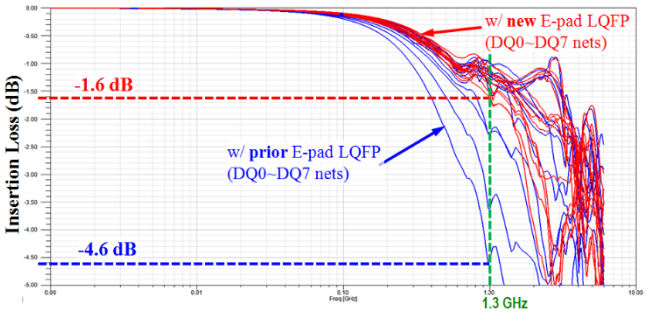


Figure 7. Insertion loss comparison of simulated channel S-parameters for DDR3 DO0-7 nets. Red curves are for the package with more bridges and blue curves are for the package with fewer bridges.

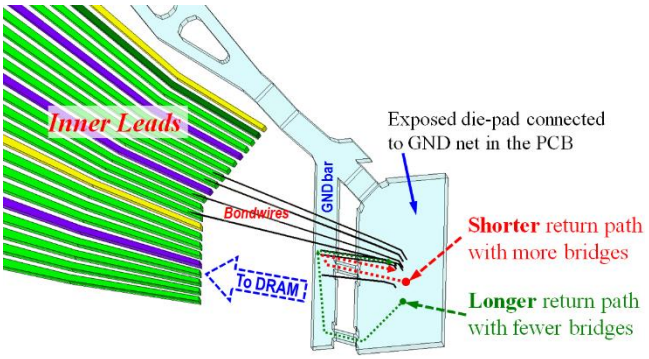


Figure 8. Return paths of high-speed signals in E-pad LQFP package.

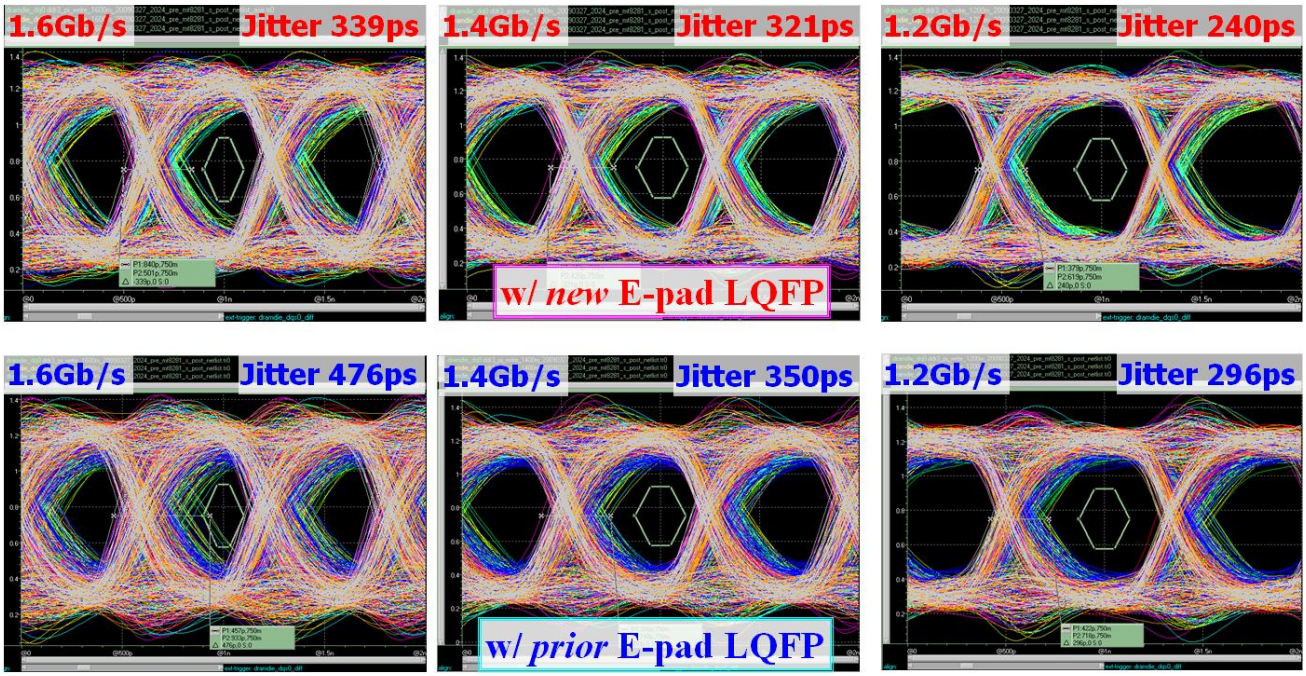


Figure 9. Simulated DDR3 eye-diagrams of overlapping 1-byte signals on DRAM chip side for writing data at 1.2, 1.4, and 1.6 Gb/s respectively. The E-pad LQFP with many bridges is in the upper charts and with few bridges is in the lower charts.

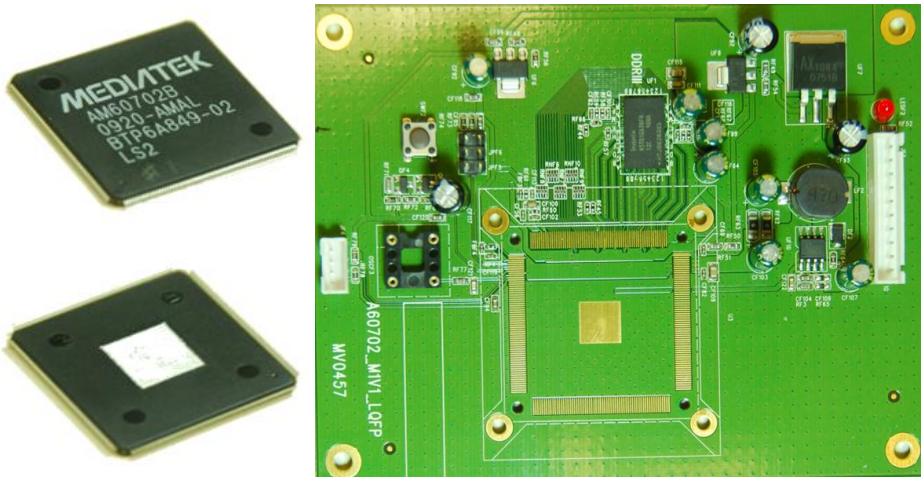


Figure 10. The practical E-pad LQFP package and 2-layer test board for DDR3 waveform verification.

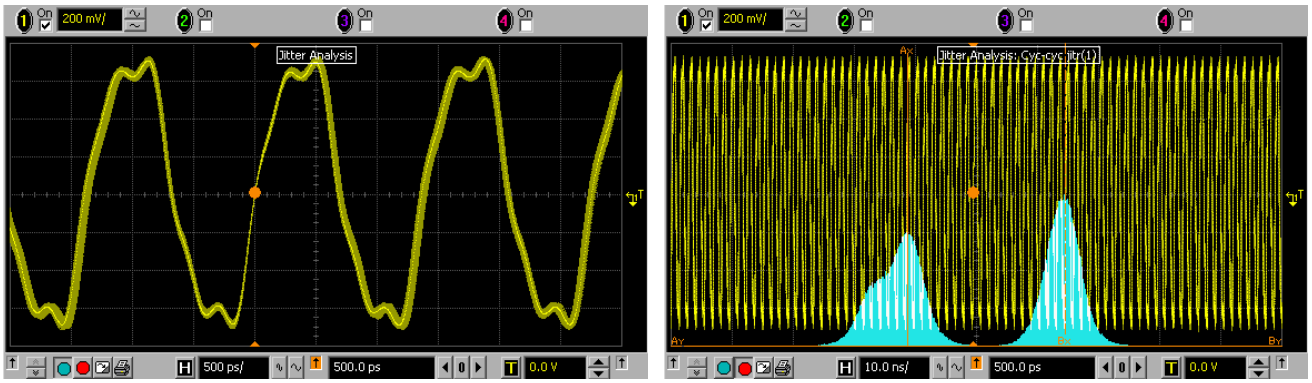


Figure 11. Measured DDR3 differential clock waveform at 729 MHz without data access (idle state) for 50.5% pulse width and tJIT (cc) for -59.4 to 56.9 ps.

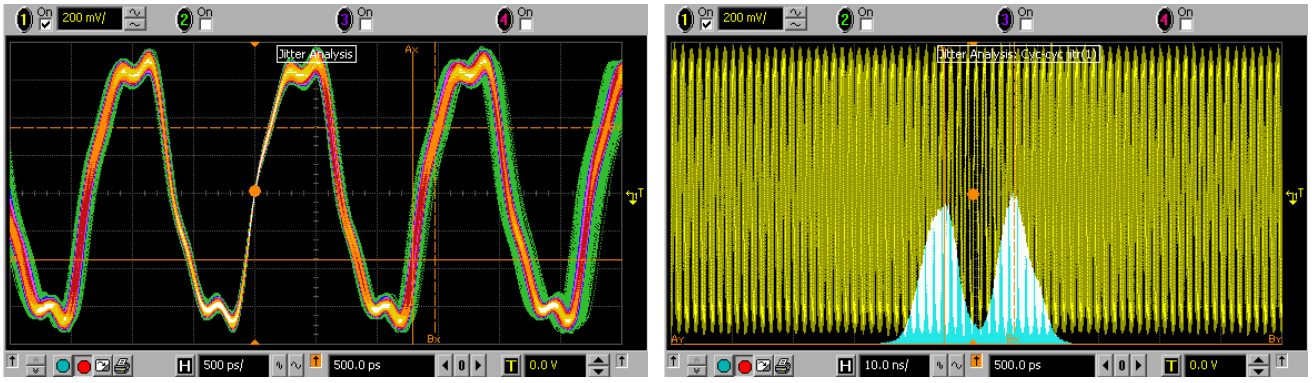
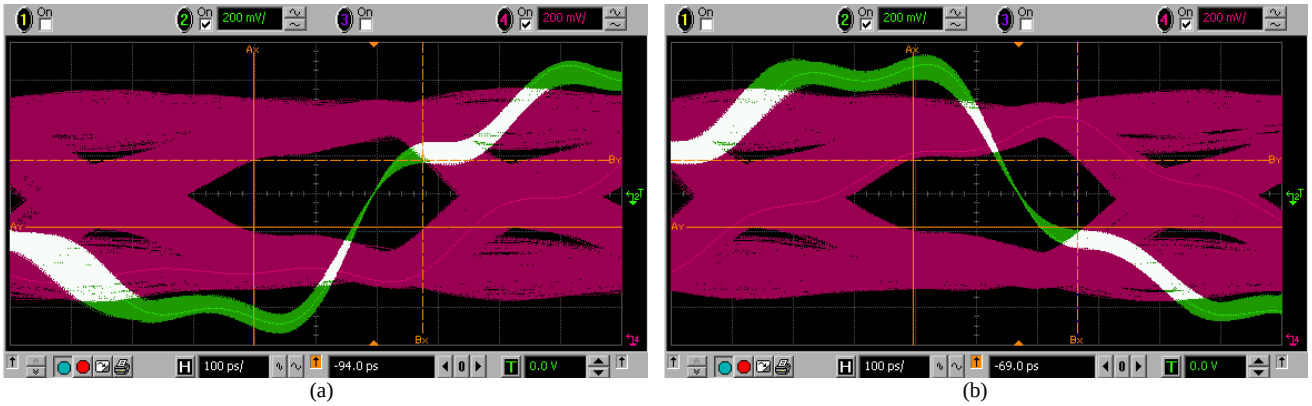


Figure 12. Measured DDR3 differential clock waveform at 729 MHz with specified data access (special test patterns) for 50.5% pulse width and tJIT (cc) for -93.0 to 90.1 ps.



(a) (b)

Figure 13. Measured DDR3 DQ14 writing waveform at 1458 Mb/s with 65 Ω drive strength for timing window. (a) 276 ps triggered by rising DQS. (b) 270 ps triggered by falling DQS.

or path is, as illustrated in Figure 8. Accordingly, the smaller the wire loop or area, the smaller the wire inductance or impedance is. That is due to high-speed or high frequency return currents follow the path of least inductance. The longer the return path, the more high-frequency components filtered out it will slow the edge rate (Johnson, Graham, 1993; Hall *et al*, 2000; Young, 2001).

Less insertion loss of full DDR3 channel in the frequency domain would expect larger eye-open in the time domain. Figure 9 demonstrates the compared DDR3 eye-diagrams between before and after package modification. Improved timing window is obviously for the new E-pad LQFP, especially for the data rate at 1.6 Gb/s. Reduced data jitter or skew achieves the larger timing margin due to less insertion loss causing less edge rate degradation. Based on the acceptable co-simulation results, there was more confident to assemble the real chip for the following verification. Leadframe packages can be manufactured using the stamping or etching process. Therefore, increase of bridge number in the package is without any cost impact.

6 DDR3 Waveform Verification

The memory controller chip was made using tsmc 40-nm process node, assembled with the modified E-pad LQFP256 package and mounted on the 2-layer test board, as shown in Figure 10. The test conditions are as follows:

1) PCB: 2 layers, 1.6 mm thickness; signal trace width/space = 5/20 mils.
2) Power supplies: 1.05 V for the core power (V_{CCK}) and 1.5 V for the I/O power (V_{CCIO}).
3) DRAM: Hynix DDR3-1600 1Gb (x16, H5TQ1G63BFR), FBGA96 package.
4) Access rates: Clock/DQS at 729 MHz and DQ/DM at 1458 Mb/s.
5) Clock (parallel) termination: 100 Ω near the DRAM.
6) Interface settings: 60 Ω ODT and 40 Ω drive strength for DRAM; 120 Ω ODT and 65 Ω drive strength for memory controller.

Figure 11 shows the differential clock waveform measured on the parallel termination (100 Ω) when there is no data access (idle state). The measured pulse width (tCH) and cycle to cycle period jitter (tJIT, cc) are 50.5% and -59.4 to 56.9 ps, respectively. When the data access with the special test patterns, the measured tJIT becomes worse, as shown in Figure 12, when all high-speed data nets start to switch resulting in larger voltage droop on the I/O power due to simultaneous switching noise (SSN). Although adding more board capacitors would stabilize the I/O power, the result was insignificant due to the high power impedance with limited power leads assigned in the E-pad LQFP256 package. The writing data (DQ14) waveform was measured near the DRAM on PCB without significant overshoot and undershoot as shown in Figure 13. This phenomenon is same with the co-simulation predicted in Figure 9. The measured timing window (setup + hold time) is around 270 ps. Eventually, All the measurement data conform to JEDEC Standard (JEDEC Std., 2012).

7 Conclusions

The co-simulation flow using the reliable simulation tools to predict the high-speed channel performance fast is presented. The modified E-pad LQFP256 package with more bridges was proposed resulting in shorter return path and achieved better timing window in the low-cost high-speed channel. All the measured waveforms meet JEDEC specification. In 2010, we rolled out the first digital TV SoC encapsulated in the E-pad LQFP256 package accessing DDR3 data rate over 1.3 Gb/s on the 2-layer PCB in the world. The next challenge is to study if the DDR4 channel could be implemented with the leadframe package.

Acknowledgements

The authors would like to thank BoWei Hsieh of MediaTek Inc., Taiwan for his helpful provision of the measured DDR3 waveforms.

References

N. Chen. *High-Speed Digital Interface Integration in Consumer Electronics–From Package to Motherboard Design*. Ph.D. diss., National Chung Hsing University, Taiwan, 2009.

Y. H. Choi. *Lead frame used for the fabrication of semiconductor packages and semiconductor package fabricated using the same*. U.S. Pat. 6437427, Aug. 20, 2002.

S. H. Hall, G. W. Hall, and J. A. McCall. *High-Speed Digital System Design*. New York. NY: John Wiley & Sons, 2000, ch. 6.

DDR3 SDRAM Specification. JEDEC Std. JESD79-3F, 2012.

H. Johnson, and M. Graham. *High-Speed Digital Design*. Upper Saddle River, NJ: Prentice-Hall, 1993, ch. 5.

Design Guide for Two DDR3-1066 UDIMM Systems. Boise, ID: Micron Technology, 2009.

J. Shah. *Improving DDR Performance by Switching from Wirebond to Flip Chip*. Oct. 10, 2012. [Online]. Available: http://embedded-computing.com/

Meeting Timing Budgets for DDR Memory Interfaces. Mountain View, CA: Synopsys, Inc., Apr. 2009.

DDR3 Design Requirements for KeyStone Devices. Dallas, TX: Texas Instruments, 2014.

B. Young. *Digital Signal Integrity*. Upper Saddle River, NJ: Prentice-Hall, 2001, ch. 5.

Automatic Generation of Dynamic Simulation Models based on Standard Engineering Data

Niklas Paganus[1] Marko Luukkainen[2] Karri Honkoila[1] Tommi Karhela[2]

[1]Fortum Power and Heat Ltd., Finland, {niklas.paganus,karri.honkoila}@fortum.com
[2]VTT Technical Research Centre of Finland Ltd., Finland, {marko.luukkainen,tommi.karhela}@vtt.fi

Abstract

Dynamic process simulation is used to mitigate risks, reduce costs and improve quality of design in plant engineering. Traditionally, simulation models are created manually from engineering source data. Benefits of utilising simulation are recognised by the industry, but simulation is not exploited to its full potential due to its current laborious nature. Engineering software interoperability improves efficiency in engineering workflows. Required manual work is reduced, enabling faster and more robust design to be conducted. In this paper, work conducted by authors in integrating the dynamic process simulation software Apros into the engineering workflow by automatically creating simulation models based on standard engineering data is reported. A case study was conducted to demonstrate the implemented features. Process engineering data in the Proteus XML format was used as the source data for simulation model generation. The case study shows that the implemented features reduce manual work required, lowering the threshold for utilising simulation.

Keywords: simulation, engineering workflow, interoperability, virtual plant

1 Introduction

Efficiency of engineering actions in plant design is wanted to be improved in process industry and power generation. Profitable and safe operations are wanted to be achieved faster. Advanced computer aided engineering (CAE) tools are utilised in the engineering workflow, allowing engineers to conduct their work efficiently. At the same time engineering projects struggle with delays and costs related to correction of design flaws recognised late in the engineering workflow or during the operation of a plant.

Two alternatives for solving these issues are discussed in this paper. First, a wider and earlier utilisation of dynamic process simulation can help in identifying engineering errors earlier, reducing or avoiding costs of corrective actions. Second, engineering software interoperability enables engineers from different disciplines to work in a more integrated manner, hence improving communication in the workflow. Thereby, time and costs required for engineering are reduced and sources for engineering errors are prevented. Additionally in this paper, it is shown how engineering software interoperability lowers the threshold for utilising dynamic process simulation by enabling automatic generation of simulation models.

Plant engineering actions are typically organised as an engineering workflow. Practices in the industry vary, but in common practises the actions are organised into design phases in which different engineering disciplines conduct their design effort. The focus disciplines of this paper are process, automation and simulation engineering.

Process engineers conduct their design effort using CAE software, resulting in piping and instrumentation diagrams (P&ID) and 3D models describing the plant. Typically, the basic design is delivered as P&IDs and the detailed design is implemented with 3D models. Typical engineering software use proprietary data models and file formats for representing design data instead of exploiting standards. This also applies to offered interfaces available in software. This restricts software interoperability and thereby both resource intensive and error-prone manual information transfer is taking place in the engineering workflow (Karhela et al, 2012; Estevez et al, 2012). However, an emerging trend of utilising standard data models and formats is emerging (Estevez et al, 2012), enabling engineering software interoperability. The ISO 15926 standard (International Organization for Standardization, 2004) and the related Proteus XML specification (Fiatech and POSC Caesar Association, 2016) are examples of standardisation being adopted by the industry. Such data formats serve as proper initial data also for extended use cases, e.g. for generating simulation models as exploited in this paper.

Simulation is utilised in the engineering workflow to answer engineering and operational questions with lower risks compared to traditional testing (Oppelt et al, 2015a). The dynamic behaviour of a plant can be analysed using dynamic process simulation. Simulation is not utilised to its full capability due to its laborious modelling requirement, partly as a result of non-interoperable engineering software (Karhela et al, 2012; Oppelt et al, 2014). Modelling effort can be significantly

reduced by automating simulation model generation based on available design data (Karhela et al, 2012). In practise this means that simulation software must be made interoperable with plant engineering software. This can be achieved by implementing customised software interfaces, but committing to standards makes the interfaces reusable. Highest efficiency can be achieved if simulation models can be generated based on the same standard engineering data as CAE software use for standard engineering data transfer.

In this paper the effort made by the authors in developing automatic simulation model generation based on standard engineering data is reported. The feature has been developed for dynamic process simulation software Apros 6, which utilises the Simantics platform. In the case study reported in this paper, a simulation model is created based on P&ID source data. The P&ID is drawn in Intergraph SmartPlant PID and exported into the ISO 15926-based Proteus XML format. An Apros model was generated based on a predefined ruleset.

The paper is structured in the following way. After this introductory section, relevant related research is presented in section 2. In section 3, the work conducted by the authors for Apros simulation software and the Simantics platform is reported. In section 4, a small case study conducted for this paper to demonstrate the implemented features is reported. Finally, the paper is concluded in section 5.

2 Related work

In this section, previous research relevant for this paper is reviewed and necessary terminology is defined. Major topics are the plant engineering workflow, engineering software interoperability and the role of simulation in plant engineering with focus on simulation model generation.

In this paper the concept plant engineering workflow is used to describe the organisation of engineering actions in a typical plant construction or retrofitting project. The concept is simplified from actual workflows in use since industry practises vary significantly. Thereby, a generalised workflow is formulated. Similar concepts used in literature are e.g. the plant engineering process (Hoyer et al, 2005), the design process (Towler et al, 2013) and the lifecycle of a process plant (Oppelt et al, 2015a). In the workflow, engineering actions are divided into design phases and participating engineering disciplines. The generalised workflow used in this paper is assumed to consist of five phases. These phases are 1. Conceptual design, 2. Basic design, 3. Detailed design, 4. Commissioning, 5. Operation and maintenance. Most relevant phases for this paper are the basic and detailed design phases.

The engineering workflow requires seamless co-operation between engineering disciplines. Engineers are accustomed to using best-in-class engineering software and interoperability has been restricted by lacking interfaces of software tools. Design data is handed over both within and across disciplines as the workflow proceeds. Design information is lost in the transfers when non-interoperable engineering software is used and significant manual action is required. The emerging trend of engineering software interoperability and utilisation of standards for representing design data will improve work conducted in the workflow as transferring design data can be automated. This requires that engineering software tools are equipped with standard interfaces to allow engineers to continue using their preferred tools.

A few possible alternatives for standard engineering data representation exist. The ISO 15926 standard (International Organization for Standardization, 2004) is being adopted by the process industry and in power plant engineering as a neutral data format. Originally, the standard did not define an actual data transfer format and for this purpose the XMpLant (Nextspace, 2015) Extensive Markup Language (XML)-based data format was developed for ISO 15926 data and was adopted widely in the industry. The XMpLant schema was made public and afterwards it has been developed by Fiatech (Fiatech, 2009) under the Proteus name. Thereby, this format is called Proteus XML. In a more recent part 8 of the ISO 15926 standard, a Web Ontology language (OWL)-based data format is defined, but it has not gained wide acceptance in industry. The Proteus XML-format allows representation of both P&ID and 3D data and contains both the geometry and attribute information in the same XML-file. Several commercial CAE tools aimed for P&ID and 3D design support Proteus or the related XMpLant format. Proteus is being actively developed by Fiatech in its IIMM project (Fiatech, 2015) and in other standardisation initiatives such as the DEXPI (DEXPI, 2016).

CAEX (Computer Aided Engineering Exchange) defined in IEC 62424 (International Electrotechnical Commission, 2008) is an engineering data exchange format that can be used for vendor independent data exchange, e.g. for P&IDs. CAEX has been used for generating simulation models (Hoernicke et al, 2015; Barth et al, 2009) and is also utilised by AutomationML. (Holm et al, 2012) The format has not been adopted by the industry as widely as ISO 15926 and it lacks an established industrial reference data library similar to the one available for ISO 15926.

ISO 10303 (International Organization for Standardization, 1994), or commonly known as STEP (Standard for the Exchange of Product model data) has been used to integrate manufacturing systems in several industries (Tursi et al, 2009), but it has not been adopted in process industry and power generation. ISO 15926 can be considered as a successor to STEP in process industries (Wiesner et al, 2011).

IFC (Industry Foundation Classes), also published as ISO 16739 standard (International Organization for Standardization, 2013), is a neutral data format maintained by buildingSMART International (buildingSMART International, 2016). It has been widely adopted in construction industry. IFC can be used to represent e.g. technical systems in buildings, such as heating and water piping, but it is not aimed for representing industrial processes and lacks component libraries for such processes. Therefore, ISO 15926 currently outrules IFC when considering needs for industrial simulation model generation. IFC is though prominent and could possibly be used partly as source data in 3D and future extensions might improve usability for industrial processes.

Dynamic process simulation is utilised in the engineering workflow to answer questions related to the dynamic behaviour of a plant. Conditions too costly or dangerous to test with traditional testing methods can be tested. In industry, despite recognised benefits of utilising simulation, it is commonly exploited only at certain times when needed and usually very late in the engineering workflow (Karhela et al, 2012). Corrective actions are more expensive and cause more disturbances to project schedules compared to if errors would have been discovered and corrected in an earlier phase (Oppelt et al, 2015b; Barth et al, 2013). Therefore, ways to support utilisation of simulation more extensively and earlier in the engineering workflow are needed to promote utilisation of simulation and gaining corresponding benefits.

Simulation models have traditionally been modelled manually based on process design data such as P&IDs and isometric drawings as printouts and by manually inspecting data from 3D models. Automation functionalities are added to the model based on automation diagrams. Simulation modelling therefore consists of combining heterogeneous data (Barth et al, 2013). Modelling work required for creating simulation models is considered laborious (Karhela et al, 2012) and research related to automating simulation model generation is actively conducted. See e.g. (Hoyer et al, 2005), (Hoernicke et al, 2015), (Barth et al, 2009), (Barth et al, 2013) and (Laakso et al, 2013) for previous work in automatic simulation model generation. P&IDs and other design documents are nowadays intelligent since CAE tools exploit object oriented features. Therefore every component on a P&ID has both its graphical appearance and attribute data defined (Barth et al, 2013), serving as a proper source for simulation model generation. By utilising automatic simulation model generation the models can be created faster, more accurate and the method is also less error-prone than manual modelling. These benefits should be enough to improve the profitability of simulation and thereby lowering the threshold for utilising simulation. This can enable simulation to be become a more integrated part of every phase of the engineering workflow.

Many previous proposals and implementations for automatic simulation model generation have been created by interfacing two specific software resulting in a custom integration. Tool specific approaches have also been implemented for Apros previously (Laakso et al, 2013; Paljakka et al, 2009). Standard interfaces allow implementation of reusable features. Standards being adopted in plant design should also be utilised when creating simulation models. The ISO 15926-based Proteus XML-format is promising as source data for automatic simulation model generation since it allows representation of both P&ID and 3D data, describes both the geometry and attribute information and is supported by major CAE software.

3 Implementation

The authors have implemented features for automatic simulation model generation based on Proteus XML data in Apros and Simantics software environment. First, the software environment is presented followed by the description of the implemented features.

Apros (Fortum and VTT Technical Research Centre of Finland, 2016) is a simulation software for modelling and simulation of dynamic processes developed and offered by Fortum and VTT Technical Research Centre of Finland since 1986. Apros has mostly been utilised in modelling and simulation of nuclear and combustion power plants with new application areas emerging. The core feature of Apros is the thermal hydraulic solver for one-dimensional two-phase flow (Porkholm et al, 2016). Apros offers an extensive library of process, automation and electrical components for modelling industrial processes. Recently, Apros has been integrated with automation engineering by introducing features for transferring automation design data in standard format to detailed design from Apros (Paganus et al, 2016). The newest version, Apros 6, is based on the Simantics platform.

Simantics (Simantics, 2016) is an open ontology-based integration platform for modelling and simulation (Karhela et al, 2012). It is managed by the THTH Association of Decentralized Information Management for Industry (THTH Association of Decentralized Information Management for Industry, 2016). Simantics offers a semantic triplestore database and a user interface based on Eclipse (The Eclipse Foundation, 2016) technology for its products. In the Apros case, the Apros solver is connected to Simantics, creating a modelling and simulation environment. A dedicated language for manipulating the Simantics database and its plug-ins has been developed, named Simantics Constraint Language (SCL) (Karhela et al, 2012). SCL can be used to perform model transformations, which

has been utilised by the authors for generating Apros simulation models.

The main objective of the work conducted by the authors was to achieve process simulation integration by exploiting standards for maximising usability in industrial projects, were a wide range of different software and practises are present. As the main source data for automatic generation of Apros models, P&IDs and 3D models were used. 3D undoubtedly gives more accurate description of the process than P&IDs and is the source for detailed accuracy in a simulation model. However, as simulation is wanted to be utilised in an earlier design phase, authors have designed a workflow where the initial simulation model is generated based on P&IDs. The initial model is made more detailed from 3D data when the source data is available.

The authors have developed a toolset for the Simantics platform for handling Proteus P&ID and 3D data in XML-format. The data import is based on automatic conversion of XML schemas to Simantics Layer0-based ontologies. Layer0 is the ontology description language used by Simantics (Karhela et al, 2012). Schema conversion tool converts XML element, complex type, and simple type definitions to Layer0 types and creates type specific relations based on XML indicators. For file import purposes, we automatically generate Java classes for SAX-based (SAX, 2004) XML parser, which processes Proteus XML files and creates respective instances into Simantics database.

We developed both P&ID and 3D visualisations of Proteus data. Proteus format uses STEP (ISO 10303) (International Organization for Standardization, 1994) standard's graphical definitions and Proteus files contain full graphical representation of the original design. Proteus P&IDs use only few graphical primitives; lines, circles, ellipses, arcs of the latter two, and text fields, making visualising P&IDs straightforward using Simantics diagramming component and Java2D (Oracle, 2016) graphics layer. 3D visualisation uses the Visualization Toolkit (VTK, 2016) and OpenCASCADE (Open CASCADE Technology, 2016). The P&ID visualisation is also used for defining the scope of the data to be transformed into Apros models. The selection is done by painting parts of the diagram, similarly to any raster graphics editor software including freehand drawing tool.

From the P&ID or 3D data imported into Simantics, Apros simulation models can be generated based on a predefined mapping ruleset from the selections made in the Proteus visualisation. The ruleset was implemented using SCL transformations, since the process is a model transformation were the Proteus data model is transformed into the Apros data model. Mappings can be of one to one, one to many or many to one type. The SCL transformation framework supports all of these mapping options. The basic feature is that equipment and components available in the Proteus data are mapped to Apros components. Attribute data available in the source data is utilised when the simulation model is parametrised. Also, different attributes needed for Apros can be calculated from the process data. Calculations are particularly relevant when an accurate 3D model is available and e.g. piping geometry can be used for calculating parameters necessary for a detailed simulation model. For attributes, the units used are also mapped to be suitable for Apros. The mappings can be altered by the user for specific needs.

The ruleset for generating Apros models has been implemented in a two-level architecture. The base of the ruleset is the reusable general features that handle the common features in Proteus, e.g. connections between components. In each separate project, an additional ruleset is defined according to the needs of the engineering project. This additional ruleset is designed to be very easy and fast to implement: it only requires definition of the corresponding component or attribute type in Proteus and Apros. By implementing this solution, the effort for taking automatic simulation model generation into use is minimised while still allowing project specific modifications. After the initial definition, the ruleset is reusable within the project.

When design and P&IDs are updated, the changes must be reflected to the simulation model. The challenge is that the user is usually required to make manual changes to the automatically generated model because of missing data and simulator specific needs. Those changes shall not be overridden. Our initial implementation used two way comparison, automatically generated simulation model with user made changes, and simulation model generated from updated P&IDs. This model does not allow automatic detection of user made changes, so to fix the situation, we have decided to use three-way comparison of the simulation model, one version as the originally generated model, one with user made changes, and one generated from updated P&IDs. Simantics database versioning capabilities (Karhela et al, 2012) allows for detecting changes that the user has made to the simulation model after it has been generated for the first time. Hence we do not need to store the original model and the model with user made changes separately. This reduces amount of data needed to be stored in the database.

Achievable accuracy of automatically generated simulation models depends heavily on the quality of the source data. During implementation it was concluded that the quality of the Proteus XML data produced by different software varies significantly. This partly relates to the constant development of the Proteus schema, creating a challenge for software vendors to keep up the pace. One example of this is the support for instrumentation. It is evident that the instrumentation support is not yet mature for full-scale industrial projects in most software tools, although basic features

are covered. The instrumentation model is being revised by Fiatech and DEXPI, which will also require revisions by software vendors.

4 Case study

A case study was conducted for this paper to test and validate the Proteus features developed for Apros and Simantics described in section 3. The scope of the study was to generate an Apros simulation model based on a simple P&ID. The P&ID was drawn in Intergraph SmartPlant PID software and then exported into the Proteus XML format. After importing the Proteus data into Simantics, an Apros simulation model was generated.

The example process in this paper is a simple pump-tank model. It consists of a water filled tank, TA-113, which is fed with water by activating the pump PU-112. The liquid level in the tank is controlled by adjusting the outward flow with a control valve in the outflow pipeline from the tank. This control requirement has also been drawn into the P&ID. The P&ID is illustrated in Figure 1.

The P&ID in Proteus XML format was imported into Simantics for inspection. The diagram was visualised correctly and the structure and attributes of the diagram can be reviewed by the user. By using the Proteus selection tool in Simantics, the scope intended for Apros model generation was chosen. In this case, the whole diagram is of interest and therefore the full content of the diagram was selected for model generation.

Mapping rules were defined to describe the correspondence between the source P&ID and Apros model. The general Proteus rule set, described in section 3, was used as a base and extended according to the specific needs of the case. The specific rules map components, attributes and specify how the unique identifiers are defined in the source data. Some simplifications were made, e.g. the impulse line and its valves have been modelled with a single level measurement component.

Apros model generation is automatic after the mapping rules have been defined. The generated simulation model is illustrated in Figure 2. Both the layout of the diagram and the data was kept consistent and attributes were transformed correctly. After model creation and setting necessary additional parameters for the model, dynamic simulation with initial values can be started. The model generated behaved as expected and controller tuning was successfully conducted. Thereby, the model was equivalent to a manually modelled model, but modelling effort was reduced. The benefit is that the model structure and initial parametrisation can be achieved faster and with fewer errors compared to manual modelling. Drawbacks of the method are that occasionally component and pipeline placements taken from the process data might not be optimal for the simulation model and some components might need to be moved after model generation. Also, many detailed modelling tasks are hard to automate and the degree of automation in model generation should be properly decided.

5 Conclusions

Benefits of utilising simulation are recognised by the industry but due to its laborious modelling effort required, dynamic process simulation is not utilised to its full potential in the engineering workflow. Laborious and error-prone manual work can be effectively reduced by automating creation of simulation models. This lowers the threshold for utilising simulation, making it

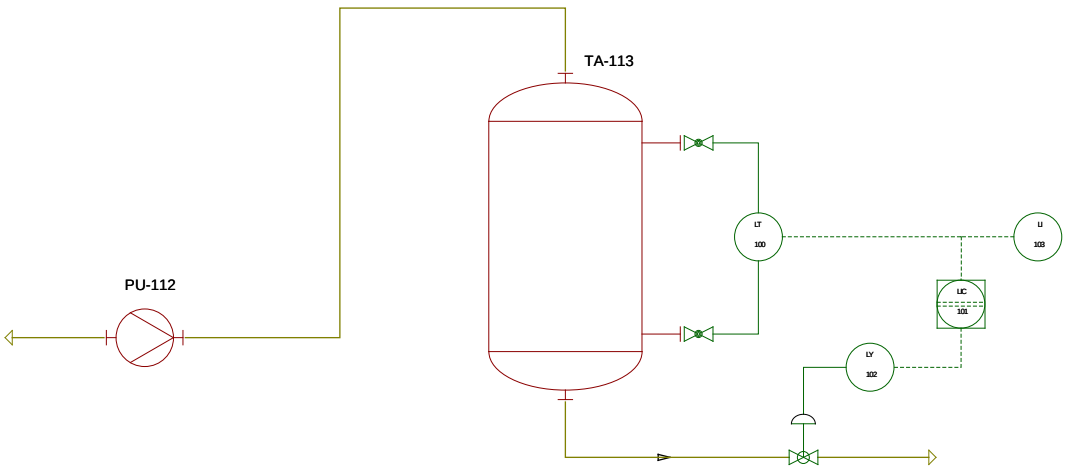


Figure 1. P&ID of the case study.

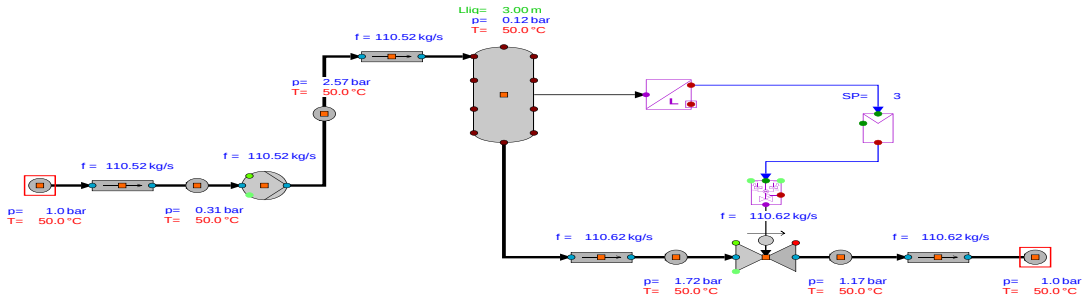


Figure 2. Apros model generated based on P&ID illustrated in Figure 1.

more profitable in plant engineering by helping in discovering design flaws in an earlier phase.

Enhanced interoperability of engineering software based on standards enable additional use cases in addition to transferring data between corresponding design tools for continued design. In this paper, one such additional usage was exploited when the work conducted by the authors in generating dynamic simulation models based on ISO 15926-based P&ID and 3D data in Proteus XML-format was reported. The features were implemented in dynamic process simulation software Apros and the Simantics platform.

The case study of the paper demonstrates successful utilisation of Proteus XML data for generating dynamic simulation models based on P&ID data. Proteus is also capable of representing 3D data, enabling creation of more accurate simulation models, which has also been tried by the authors. Quality of the source data determines the accuracy achievable for the automatically generated simulation models. Therefore, the standard interfaces in engineering software used and the engineering practises applied in the engineering workflow shall support production of high quality data. The authors have tested CAE software from different vendors and concluded that the quality of standard interfaces varies and the interfaces need to become more mature to be effectively applicable in major industrial projects.

Utilising standard engineering data in engineering workflows is promising and a wider utilisation is demanded by the industry. One risk for a wider utilisation is whether the industry and software vendors can agree on a common standard since many promising alternatives currently exist. Therefore, co-operation between standardisation organisations, the industry and software vendors is required. Currently, availability and quality of interfaces vary but work conducted by e.g. DEXPI show the interest for interoperability based on standards by industry, software vendors and academia.

Further validation is required for the implemented methods for automatic simulation model generation. The authors are going to validate the features in an extensive industrial use case to analyse the benefits and drawbacks more detailed. Simplification of simulation models by altering the nodalisation should also be handled in the automatic model generation.

References

M. Barth, M. Strube, A. Fay, P. Weber, and J. Greifeneder. Object-oriented engineering data exchange as a base for automatic generation of simulation models. In *Annu Conf. IEEE Ind. Electron. Soc.*, Porto, 2009. doi: 10.1109/IECON.2009.5415229.

M. Barth and A. Fay. Automated generation of simulation models for control code tests. *Control Eng. Practice*, 21(2): 218-230, 2013. DOI: 10.1016/j.conengprac.2012.09.022.

buildingSMART International. *IFC Overview summary*. Updated: 2016. Accessed: 18.5.2016. Available: http://www.buildingsmart-tech.org/specifications/ifc-overview.

DEXPI. *DEXPI - Data Exchange in the Process Industry*. Updated 2016. Accessed 16.3.2016. Available: http://www.dexpi.org/.

E. Estevez and M. Marcos. Model-based validation of industrial control systems. *IEEE Trans. Ind. Inform.*, 8(2): 302-310, 2012.

Fiatech. *The Proteus project*. Updated 2009. Accessed: 16.3.2016. Available: http://fiatech.org/index.php/?option=com_content&view=article&id=1115&Itemid=748.

Fiatech. *ISO 15926 Information Models and Proteus Mappings IIMM*. Updated 2015. Accessed 16.3.2016. Available: http://fiatech.org/information-

management/projects/1161-iso-1592 6-information-models-and-proteus-mappings-iimm.

Fiatech and POSC Caesar Association. *Proteus XML specification documents and XML schema*. Updated: 2016. Accessed: 18.5.2016. Available: http://fiatech.org/information-management/ projects/1161-iso-15926-information-models-and-proteusmappings-iimm.

Fortum and VTT Technical Research Centre of Finland. *Apros Process Simulation Software*. Updated: 2016. Accessed: 9.3.2016. Available: http://www.apros.fi/en/.

M. Hoernicke, A. Fay, and M. Barth. Virtual plants for brown-field projects. In *IEEE Conf. Emerging Technol. Factory Autom.*, Luxembourg, 2015. doi: 10.1109/ETFA.2015.7301462.

T. Holm, L. Christiansen, M. Goring, T. Jager, and A. Fay. ISO 15926 vs. IEC 62424 — Comparison of plant structure modeling concepts. In *IEEE Conf. Emerging Technol. Factory Autom.*, Krakow, 2012. doi:10.1109/ETFA.2012.6489662.

M. Hoyer, R. Schumann, and G. C. Premier. An approach for integrating process and control simulation into the plant engineering process. *Comput. Aided Chemical Eng.*, 20(B): 1603-1608, 2005. doi: 10.1016/S1570-7946(05)80109-9.

International Electrotechnical Commission. IEC 62424:2008. *Representation of process control engineering – Requests in P&I diagrams and data exchange between P&ID tools and PCE-CAE tools*. 1st ed. 2008.

International Organization for Standardization. ISO 10303-1:1994. *Industrial automation systems and integration – Product data representation and exchange – Part 1: Overview and fundamental principles*. 1st ed. 1994.

International Organization for Standardization. ISO 15926-1:2004. Industrial automation systems and integration – *Integration of life-cycle data for process plants including oil and gas production facilities – Part 1: Overview and fundamental principles*. 1st ed. 2004.

International Organization for Standardization. ISO 16739:2013. *Industry Foundation Classes IFC for data sharing in the construction and facility management industries*. 1st ed. 2013.

T. Karhela, A. Villberg, and H. Niemistö. Open ontology-based integration platform for modeling and simulation in engineering. *Int. J. Modeling, Simulation, and Scientific Computing*, 3(2): 1250004-1-1250004-36, 2012.

P. Laakso, J. Lappalainen, T. Karhela, and M. Luukkainen. Virtual plant combines engineering tools for the process industry. VTT Research Highlights 8, 2013. Accessed: 11.5.2016. Available: http://www.vtt.fi/inf/pdf/researchhighlights/ 2013/R8.pdf.

Nextspace. *XMpLant*. Updated 2015. Accessed: 16.12.2015. Available: http://www.nextspace.co.nz/products-and-services/solutions/xmplant/.

Open CASCADE. *Open CASCADE Technology*. Updated: 2016. Accessed 16.5.2016. Available: http://www.opencascade.com/.

M. Oppelt and L. Urbas. Integrated virtual commissioning an essential activity in the automation engineering process: From virtual commissioning to simulation supported engineering. In *Annu. Conf. IEEE Ind. Electron. Soc.*, Dallas, TX, 2014. doi: 10.1109/IECON.2014.7048867.

M. Oppelt, M. Barth, and L. Urbas. The Role of Simulation within the Life-Cycle of a Process Plant. Results of an online survey, 2015a. doi: 10.13140/2.1.2620.7523.

M. Oppelt, G. Wolf, and L. Urbas. Towards an integrated use of simulation within the life-cycle of a process plant. In *IEEE Conf. Emerging Technol. Factory Autom.*, Luxembourg, 2015b. doi: 10.1109/ETFA.2015.7301521.

Oracle. *Java 2D Graphics and Imaging*. Updated: 2016. Accessed: 18.5.2016. Available: https://docs.oracle.com/javase/8/docs/techn otes/guides/2d/.

N. Paganus, K. Honkoila, and T. Karhela. Integrating dynamic process simulation into detailed automation engineering. In *21st IEEE International Conference on Emerging Technology and Factory Automation ETFA'2016*, 6-9.9.2016, Berlin, Germany .

M. Paljakka, J. Talsi, and H. Olia. Experiences on the integration of automation CAE and process simulation tools - case FupRos. In *Automaatio XVIII Seminar, Finnish Society of Automation, publication series 36*, Helsinki, Finland, 2009.

K. Porkholm, H. Kontio, H. Plit, M. Mustonen, and K. Söderholm. APROS simulation model for Olkiluoto-3 EPR Applications. In *European Nuclear Society TopSafe Trans.*, Dubrovnik, Croatia, 2008. Accessed: 9.3.2016. Available: http://www.euronuclear.org/events/topsafe/transactions/TopSafe2008-transactions-poster.pdf.

SAX. *SAX – Simple API for XML*. Updated: 2004. Accessed: 18.5.2016. Available: http://sax.sourceforge.net/.

Simantics. *Open operating system for modeling and simulation*. Updated: 2016. Accessed: 14.4.2016. Available: https://www.simantics.org/.

Eclipse. *The Eclipse Foundation open source community website*. Updated 2016. Accessed 16.5.2016. Available: http://www.eclipse.org/.

THTH. *THTH Association of Decentralized Information Management for Industry*. Updated: 2016. Accessed: 18.5.2016. Available: http://www.ththry.org/?lang=en.

G. P. Towler and R. K. Sinnott. *Chemical Engineering Design: Principles, Practice, and Economics of Plant and Process Design*. 2nd ed. Oxford, UK, Butterworth-Heinemann, 2013.

A. Tursi, H. Panetto, G. Morel, and M. Dassisti. Ontological approach for products-centric information system interoperability in networked manufacturing enterprises. *Annual Reviews in Control*, 33(2): 238-245, 2009. DOI: 10.1016/j.arcontrol.2009.05.003.

VTK. *VTK - The Visualization Toolkit*. Updated: 2016. Accessed 16.5.2016. Available: http://www.vtk.org/.

A. Wiesner, J. Morbach, and W. Marquardt. Information integration in chemical process engineering based on semantic technologies. *Computers & Chemical Engineering*, 35(4): 692-708, 2011. DOI: 10.1016/j.compchemeng.2010.12.003.

A Novel Credibility Quantification Method for Welch's Periodogram Analysis Result in Model Validation

Yuchen Zhou, Ke Fang, Kaibin Zhao, Ping Ma*

Control and Simulation Center, Harbin Institute of Technology, Harbin, P.R. China
ZhouYuChen-01@163.com, FangKe@hit.edu.cn, ZhaoKaiBin1986@163.com, PingMa@hit.edu.cn

Abstract

Welch's periodogram is widely used in frequency domain model validation. However, Welch's analysis results just reveals whether the time series passed the consistency test in each discrete frequency point, which is not a quantitative credibility evaluation result and may not help the evaluation expert to grade credibility level of simulation system. Based on Welch's periodogram and consistency test approach, a novel credibility quantification method using weight density function is proposed. Furthermore, the frequency analysis and credibility quantification process is provided. Finally, the credibility quantification of radiated noise in ship acoustic feature simulation indicates the method proposed is effective for periodic time series with complicated spectrum.

Keywords: Welch's periodogram, credibility quantification, line spectrum, periodic time series, model validation

1 Introduction

The M&S technology has the advantages of economy, security, repeatability and nondestructive, which makes it widely used in aerospace, nuclear, communication, et al. Verification Validation and Accreditation (VV&A) should be conducted to guarantee the validity of complex simulation system (Oberkampf 2008; Sargent, 2013; Wang, 2000). Through behavior-similarity analysis between simulation system and real-world system, we can obtain the credibility of the simulation system.

Frequency analysis method is usually used for the validation of periodic time series. Fishman and Kiviat firstly proposed frequency domain validation method and applied it to queuing model validation (Fishman, 1967). Gallanteta1 put forward a data consistency analysis approach based on Analysis of Variance (ANOVA) and periodogram method. Montgomery used spectral analysis to evaluate the credibility of missile simulation system (Montgomery, 1980&1983).

To resolve the thermal challenge problem suggested in reference (Roy, 2011), (Ferson, 2008) proposed an area metric, which takes the integral over the area difference between the cumulative distribution function(CDF) of simulation data and the empirical CDF of the measured samples as the disagreement between the simulation model and real-world system (Li, 2014; Sankararaman, 2011). Mullins classified the data scenarios with aleatory and epistemic uncertainty and studied how different validation metrics may be appropriate for varies data samples (Mullins, 2015). Literature (Zhang, 2011) provided a group AHP method to evaluate the credibility of complex simulation system, in which Hadamard convex combination is used to aggregate the judgement matrices constructed by different assessment experts.

In the frequency analysis, some spectrum of simulated data and observed data are extremely complicated. Consistency test result of spectrum just shows whether the data passed the examination in each discrete frequency point rather than a quantitative credibility evaluation result, which may not help the evaluation expert to grade the credibility level of simulation system. How to transform the consistency test result to credibility is the key problem to resolve in this paper. The credibility quantitative method is illustrated in detail and case study demonstrates the frequency analysis and credibility transform process of radiated noise comprehensively.

2 Periodogram Method

The frequency domain analysis method involves power spectrum density estimation and consistency test. Power spectral density estimation is a data transform approach to estimate the distribution of signal energy in each frequency points using limited data. Welch's periodogram is a typical method in frequency domain validation.

2.1 Welch's Periodogram Method

Periodogram method uses Fast Fourier Transform (FFT) algorithm to estimate spectrum of stationary random sequence and the estimated spectrum is sensitive to the length of time series. If the data length is beyond a threshold value, spectrum oscillated

intensely. On the contrary, if the data length is reduced to a certain extent, the resolution of spectral may decrease and the estimation error increase significantly.

Welch's spectral estimation is an improvement of periodogram by dividing the time series into many data segments and using non-rectangular windows to handle each data segments (Welch, 1967). When using Welch's method, the time series $x(n) \in R^N$ is divided into $K = \frac{N}{L}$ pieces, as:

$$x^{(i)}(n) = x(n + iL - L), 0 \le n \le L-1, 1 \le i \le K \quad (1)$$

Each data segments is including L samples, and we can calculate K modified periodogram by:

$$J_L^{(i)} = \frac{1}{LU}\left|\sum_{n=0}^{L-1} x^{(i)}(n) w(n) e^{-j\omega n}\right|^2, i = 1, 2, ..., K \quad (2)$$

where $U = \frac{1}{L}\sum_{n=0}^{L-1} w^2(n)$ denotes the mean power of window function.

Finally the power spectrum of time series $x(n)$ can be calculated by:

$$S_x^w(\omega) = \frac{1}{K}\sum_{i=1}^{K} J_L^{(i)}(\omega) \quad (3)$$

2.2 Consistency Test

Suppose that $S_x(\omega)$ and $S_y(\omega)$ are spectrum of simulation model output time series $x(n)$ and real-world system output time series $y(n)$. The estimated spectrum are $S_x(\omega)$ and $S_y(\omega)$. As is proved $\frac{rS(\omega)}{S(\omega)} \sim \chi_r^2$ (Chen, 1988), we can make the null hypothesis and alternative hypothesis as:

$$H_0 : S_x(\omega) = S_y(\omega)$$
$$H_1 : S_x(\omega) \ne S_y(\omega)$$

Statistics of hypothesis test is:

$$F = \frac{rS_x(\omega) / S_x(\omega) / r}{rS_y(\omega) / S_y(\omega) / r} \sim F(r, r) \quad (4)$$

where $r = 2N / \sum_{k=-M}^{M} w^2(k)$.

If the original hypothesis is accepted under the confidence level α, we can draw:

$$p\{F_{\frac{\alpha}{2},r,r} \le \frac{S_x(\omega)}{S_y(\omega)} \le F_{1-\frac{\alpha}{2},r,r} \mid H_0\} = 1 - \alpha \quad (5)$$

On the contrary, if $\frac{S_x(\omega)}{S_y(\omega)} \notin \left[F_{\frac{\alpha}{2},r,r}, F_{1-\frac{\alpha}{2},r,r}\right]$, we should accept the alternative hypothesis. Finally, the F test can be made in every frequency point ω_i.

3 Credibility Quantification Method

3.1 Credibility Quantification Method

Since the complexity of computational systems, we may not take the validation result under single input condition as the final credibility of the simulation system. Literature (Mullins, 2015) provides an approach to integrate the model validation results from multiple simulation scenarios, which is defined by equation (6).

$$v_{overall} = \int v(x)\pi(x)dx \quad (6)$$

where x is a n-dimensional vector of input conditions and $v(x)$ is the validation result under input condition x; $\pi(x)$ is the joint probability density of the point x. Mullins et al. propose that $\pi(x)$ is a weighting function for the importance of multiple validation results which can be estimated according to the relevance of each experiment conditions to the overall intended use of the computational model.

Provided that limited data samples are obtained, $v(x)$ is available at finite points. Therefore, equation (6) is reformulated as

$$v_{overall} = \sum_{i=1}^{m} v(x) w_i \Big/ \sum_{i=1}^{m} w_i \quad (7)$$

Literature (Zhang, 2010) studied transform algorithm of several commonly used validation methods. A direct transform approach for periodogram is provided, which takes the percentage of frequency range passed consistency check accounting for the whole frequency range as credibility. This conversion method is based on the hypothesis that statistic characteristic consistency of power spectrum in each frequency point has same effect on the overall credibility. Actually, since the power spectrum density is higher in one or several frequency bands, the effect on the similarity of the consistency test in each frequency points is not equal. Therefore, an inaccurate result may be calculated using direct transform approach.

Power spectrum density curves can be divided into three categories, broadband spectrum (Figure 1), line spectrum (Figure 2), and mixed spectrum (Figure 3) which is composed of broadband spectrum and several line spectrum. In general, the frequency bands with higher energy often reflect the periodic characteristics of the system.

The power difference of line spectrum between predicted time series and observed time series is usually the main reason for failing to pass the consistency check. Thus, the transform of line spectrum should be given high priority in the process of credibility quantification.

Based on the above analysis, the credibility quantification process should focus on the key frequency band with higher power. Therefore, we may use the normal weight density function to realize the

conversion. The integral of normal weight density function $f(\omega)$ among the whole frequency band is 1 and the multiple of weight function $f(\omega)$ and consistency test result is the credibility.

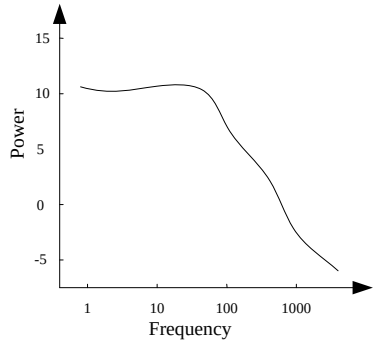


Figure 1. Broadband spectrum.

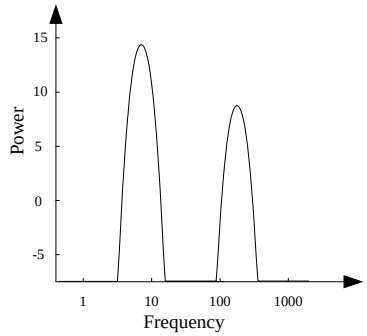


Figure 2. Line spectrum.

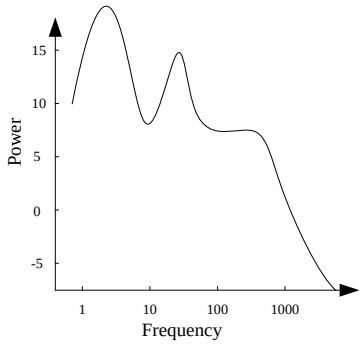


Figure 3. Mixed spectrum,

The normal weight density function is defined as:

$$f(\omega)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(\omega-\mu)^2}{2\sigma^2}),\ \omega\in[0,+\infty) \quad (8)$$

where μ is the center position of weight density function and σ is the dispersion parameter.

Let $F(\omega_0)$ be the weight integration of $f(\omega)$ in [0, ω_0], as:

$$F(\omega_0)=\int_0^{\omega_0} f(\omega)d\omega \quad (9)$$

then the weight integration of $f(\omega)$ among $q=[a,b]$ can be expressed as $F(b)-F(a)$.

Suppose the frequency points passed the consistency test as 1, otherwise as 0. The whole frequency band is divided into two sections, the band M passed the check and the band N failed to pass the check. The two sections are defined as:

$$M=\bigcup_{k=1}^{p} m_i=\bigcup_{k=1}^{p}[m_{ka},m_{kb}] \quad (10)$$

$$N=\bigcup_{k=1}^{q} n_i=\bigcup_{k=1}^{q}[n_{ka},n_{kb}] \quad (11)$$

Then choose appropriate parameters for each weight functions and the integral of band M is the final credibility.

$$C=F(M)=\sum_{k=1}^{p}F(m_k) \quad (12)$$

Finally, the parameters of normal weight density function can be chosen as follows.

1) The mean of the normal distribution should be the middle of the line spectrum.

2) 3σ Rule may be followed in the selection of the variance.

According to the parameter selection approach, the line spectrum will cover the frequency interval $[\mu-3\sigma,\mu+3\sigma]$. Although $\lim_{\omega\to+\infty}F(\omega)=\int_0^{\omega}f(\omega)d\omega$ $=1-\Delta<1$, the weight loss $\Delta<0.15\%$ and it may be ignored. For example, actuator works periodically in control system. Figure 4 is the power spectrum of actuator output time series. The work frequency of the actuator is between 8 Hz and 10Hz. Thus, we may choose $\mu=9$, $\sigma=0.67$ as the parameters of weight function (Figure 5).

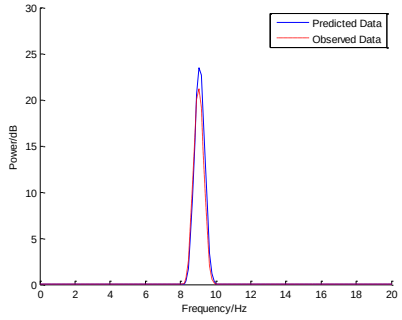


Figure 4. Sperctem of actuator output time series.

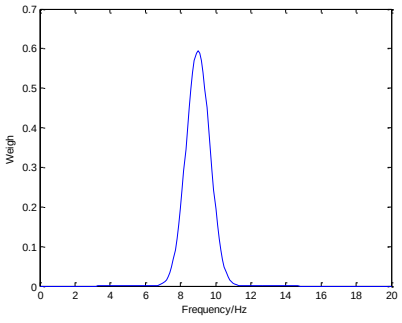


Figure 5. Normal weight density function.

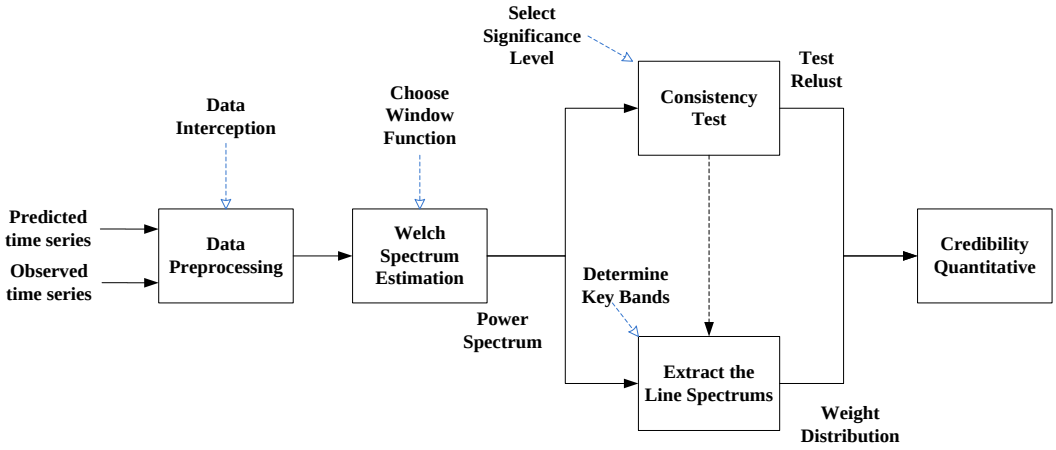


Figure 6. Frequency analysis and credibility quantification process.

3.2 Frequency Analysis and Credibility Quantification Process

Based on the above method, including welch analysis method and transform approach, the specific transform steps (Figure 6) are provided as follows.

Step 1. Preprocessing the predicted time series and observed time series, including data interception and smoothing filtering.

Step 2. Choose appropriate transformation points and estimate the frequency spectrum of predicted time series and observed time series.

Step 3. Select a significance level and make the consistency test.

Step 4. Extract line spectrums in power spectrum, and analysis the reasons for the bands failed to pass the consistency test. Based on the test result, separate the whole band into two sections (band *M* passed the consistency test and band *N* failed to pass é the consistency test).

Step 5. Select appropriate weight functions for each line spectrums and calculate the final credibility using equation (12).

Actually, the overall simulation credibility of computational model should integrals multiple validation results of simulation data and measured samples under varies input conditions. This paper focus on the credibility quantification of frequency-domain validation result under single data scenario.

4 Case study

The task of ship acoustic feature simulation is not only to study and analyze the acoustic characteristics of the ship, but also to realistically simulate the characteristics of the ship and applied it to the sonar system testing and ship type identification. In the research of ship acoustic feature simulation, to guarantee the characteristics similarity of radiated noise between reconstruction model and real ship, frequency domain validation can be conducted to evaluate the credibility of the ship radiated noise reconstruction model.

Radiated noise is mainly composed of mechanical noise, propeller noise and hydrodynamic noise. In general, radiated noise power spectrum is a typical representative of mixed spectrum. In the radiated noise spectrum, line spectrum reflects the periodic part of the energy distribution of the noise in the signal, which primary covers low frequency band. The line spectrum is the main feature to recognize ship types.

According to the frequency analysis and credibility quantification procedure, radiated noise data is processed as the following.

i) Data preprocessing

Radiated noise data preprocessing includes data interception and normalization. Figure 7 reveals the simulation data and reference data in time domain.

ii) Welch's spectrum estimation

Using Welch's method to estimate the power spectrum and draw the spectrum graphics. In Figure 8, since the spectrum power between 1 Hz~1000Hz is above 0dB, the band 1Hz~1000Hz is the key band in the following analysis. There emerge seven line spectrums in the Welch spectrum between 1Hz~1000Hz, including 10Hz~54Hz, 82Hz~180Hz, 200 Hz ~240Hz, 320~360HZ, 460Hz~510Hz, 722Hz~785Hz and 850~935Hz.

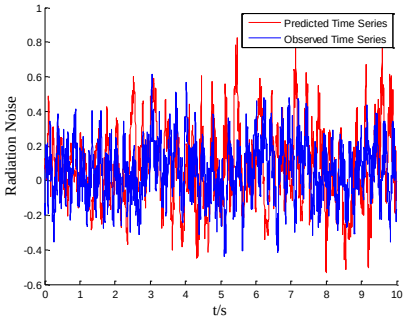


Figure 7. Comparison between predicted and observed data in time domain.

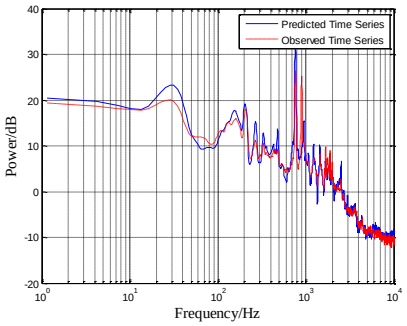


Figure 8. Power spectrum estimated by Welch's method.

Analytic hierarchy process (AHP) is used to obtain the weights of every line spectrum. According to the power and band length of each line spectrum, the judgment matrix is defined as:

$$J_D = \begin{bmatrix} 1 & 1/2 & 1 & 1/2 & 1/2 & 2 & 2 \\ 2 & 1 & 1 & 2 & 2 & 3 & 3 \\ 1 & 1 & 1 & 1 & 1 & 2 & 2 \\ 2 & 1/2 & 1 & 1 & 1 & 1 & 2 \\ 2 & 1/2 & 1 & 1 & 1 & 2 & 2 \\ 1 & 1/3 & 1 & 1 & 1/2 & 1 & 1 \\ 1 & 1/3 & 1/2 & 1/2 & 1/2 & 1 & 1 \end{bmatrix}$$

The weight vector of line spectrum is:

$$\omega = [0.122,\ 0.248,\ 0.164,\ 0.150,\ 0.150,\ 0.086,\ 0.080]$$

The coincidence index is 1.32 and the ratio is:

$$CR = CI / RI = 0.0158 < 0.1$$

Therefore the weight vector calculated is valid.

iii) Consistency test

We can use F test to check the behavioral consistency between simulation data and reference data in frequency domain. The consistency analysis result is shown in Figure 9. Under log-log coordinate system, for a frequency point, if the confidence interval contains 0dB, then the radiated noise data passed the consistency analysis on this point.

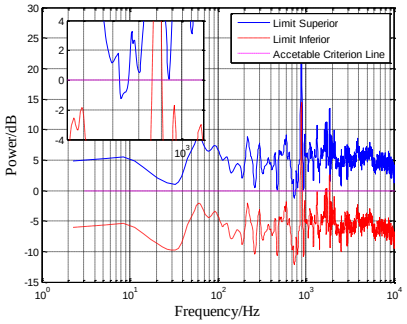


Figure 9. Consistency test result of Welch's power spectrum.

Obviously, there are two frequency bands failed to pass the check, band 736 Hz ~760 Hz and band 866 Hz ~915 Hz.

In figure 8, the power difference of line spectrum 722 Hz ~785Hz excess 5dB. This is the reason for band 736 Hz ~760 Hz failed to pass the consistency test. Meanwhile, there exists a line spectrum in 858Hz~ 931Hz for the simulated time series and another in 850Hz~935Hz for the observed time series. The frequency band difference leads to the test failure in band 866 Hz~915 Hz.

iv) Credibility quantification

According to the rule choosing weight function parameters, the parameters of seven line spectrums are defined as Table 1.

Table 1. Parameters of Normal Weight Density Function.

Function	Line spectrum	μ	σ
F_1	10Hz~54Hz	38	7.33
F_2	95Hz~180Hz	137.5	28.33
F_3	200 Hz ~240Hz	220	6.67
F_4	320~360HZ	340	6.67
F_5	462Hz~512Hz,	487	15.00
F_6	722Hz~785Hz	753.5	10.50
F_7	850~935Hz	892.5	14.17

Based on the weight of each line spectrum ω and parameters of normal weight density function, the weight on the whole frequency band is shown in Figure 10.

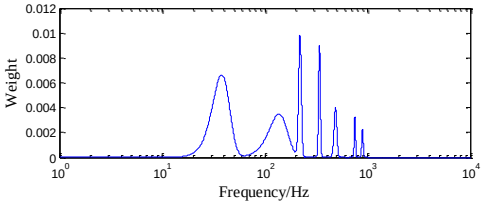


Figure 10. Weight on the whole frequency band.

Since most part of the line spectrum passed the consistency test, we can calculate the credibility loss $\sum_{k\in\Theta} L_k$ and the credibility is $1-\sum_{k\in\Theta} L_k$.

The credibility loss L_1 in band 736 Hz ~760Hz is:

$$L_1 = \left[F_6(760) - F_6(736)\right] \times 0.086 = 0.059$$

The credibility loss L_2 in band 866 Hz ~915Hz is:

$$L_2 = \left[F_7(915) - F_7(866)\right] \times 0.080 = 0.073$$

Then, the final credibility is:

$$C = 1 - L_1 - L_2 = 0.868$$

As a comparison, we use direct transform approach (Zhang J. Y, 2010) to calculate the credibility:

$$C_{dt} = \left[(736-760) + (915-866)\right]/1000 = 0.927$$

From the Welch's spectrum estimated result and consistency test result, the behavior of simulated time series and observed time series are similar to each other in most parts of the frequency band except two. The credibility quantification result confirms this

conclusion. The credibility calculated by direct transform approach is about 0.927, which is lack of factual basis. To summarize, the credibility quantification approach should focus on key line spectrums and the case study proves the quantification method proposed is effective for time series with complicated spectrum.

5 Conclusions

Based on Welch's periodogram and consistency test approach, a novel credibility quantification method using weight density function is proposed. Compared to the direct transform approach, which assume all the point have same effect on the final credibility, the credibility quantification method proposed based on normal weight density function focus on the transform of Welch's analysis result on key frequency band. The credibility quantification of ship radiated noise data proves the transform approach is reasonable. Meanwhile, the credibility quantification process indicates the method provided is effective for periodic time series with complicated spectrum.

Frequency-domain analysis in model validation is a kind of consistency test method based on pattern. Even though there exists uncertainty in simulation model input, frequency-domain approach, including Welch's analysis, Maximum Entropy Spectral Estimation (MESE) et al, can still be utilized to analyze the periodic time series. For the time series with sophisticated spectrum, extraction of line spectrum highly depends on the evaluation expert, which is time-consuming and boring for the analyst. In the future, method will be studied to obtain the accurate information of each line spectrum automatically. Furthermore, a frequency-domain validation and credibility quantification tool will be developed to improve the efficiency of model validation.

Acknowledgment

This work was supported by National Science Foundation of China (No. 61374164).

References

Z. G. Chen. *Time Series and Spectrum Analysis*. Science Press, 253-282, 1988.

S. Ferson, W. L. Oberkampf and L. Ginzburg. Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics & Engineering*. 197(29-32), 2408-2430, 2008. doi: 10.1016/j.cma.2007.07.030.

G. S. Fishman and P. J. Kiviat. The analysis of simulation-generated time series. *Management Science*, 13(7), 525-557, 1967. doi: 10.1287/mnsc.13.7.525.

W. Li, W. Chen, C. Jiang, Z. Z. Lu and Y. Liu. New validation metrics for models with multiple correlated responses. *Reliability Engineering & System Safety*, 127, 1-11, 2014. doi: 10.1016/j.ress.2014.02.002.

D. C. Montgomery and R. G. Conard. Comparison of simulation and flight-test data for missile systems. *Simulation*, 34(2), 63-72, 1980. doi: 10.1177/003754978003400206.

D. C. Montgomery and L. Greene. Methods for validating computer simulation models of missile systems. *Journal of Spacecraft and Rockets*, 20(3), 272-278, 1983. doi: 10.2514/3.25592.

J. Mullins, Y. Ling, S. Mahadevan, L. Sun and A. Strachan. Separation of aleatory and epistemic uncertainty in probabilistic model validation. *Reliability Engineering & System Safety*, 147, 49-59, 2015. doi: 10.1016/j.ress.2015.10.003.

W. L. Oberkampf and T. G. Trucano. Verification and validation benchmarks. *Nuclear Engineering and Design*, 238(3), 716-743, 2008. doi: 10.1016/j.nucengdes.2007.02.032.

C. J. Roy and W. L. Oberkampf. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, 200(25-28), 2131-2144, 2011. doi: 10.1016/j.cma.2011.03.016.

R. G. Sargent. Verification and validation of simulation models. *Journal of Simulation*, 7(1), 12–24, 2013. doi: 10.1057/jos.2012.20.

S. Sankararaman and S. Mahadevan. Model validation under epistemic uncertainty. *Reliability Engineering & System Safety*, 96(9), 1232-1241, 2011. doi: 10.1016/j.ress.2010.07.014.

Z. C. Wang. Research on simulation theory. *Journal of System Simulation*. 12(6), 604-608, 2000. doi: 10.3969/j.issn.1004-731X.2000.06.007.

P. D. Welch. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoutics*, 15(2), 17-20, 1967. doi: 10.1109/tau.1967.1161901.

J. Y. Zhang. *Validation methods and assistant tools based on coherence of data*. Master Thesis of Harbin Institute of Technology, 32-34, 2010.

Z. Zhang, K. Fang and M. Yang. Method for complex simulation credibility evaluation based on group AHP. *System Engineering and Electronics*, 33(11), 2569-2573, 2011. doi: 10.3969/j.issn.1001-506X.2011.11.42.

Identification Scheme for the Nonlinear Model of an Electro-Hydraulic Actuator

W.C. Leite Filho[1] J. Guimaraes[2]

[1] Space Mechanics and Control Division, Instituto Nacional de Pesquisas Espaciais, Brazil, waldclf@gmail.com
[2] Space Mechanics and Control Division, Instituto Nacional de Pesquisas Espaciais, Brazil, julia.guimaraes@inpe.br

Abstract

This work presents the building of a nonlinear model of an electro-hydraulic actuator in order to understand the limit cycle phenomenon that appears when it is used in a closed loop control system. Previously, a first harmonic analysis had been used to identify that system, but the results were unsatisfactory. Therefore, this work aims to build on that model with the use of Fast Fourier Transforms as a way to recognize previously unseen nonlinearities. Hardware in the loop tests are then used in order to find the proper parameters that create a particular limit cycle. Simulation results show that such approach is successful.

Keywords: nonlinear model, FFT, hardware in the loop, actuator model

1 Introduction

As part of the design of control systems of space vehicles, it is important to achieve a thorough understanding of each element modeled so that the simulated results will correctly represent the real scenario.

In particular, it is important to be able to reproduce the effect that nonlinearities create on the final output of the system, since strategies used to deal with bending modes affect the limit cycle generated by those nonlinearities.

In order to support such development, hardware in the loop (HWIL) simulations were used in an attempt to identify a proper model for the actuator, but the model proposed at the time was incomplete (Bueno and Leite Filho, 2003). A similar approach is used now in order to obtain initial values for the nonlinearities, while analysis of the Fast Fourier Transform of the signal is used to infer the missing elements.

2 Initial Configuration

For the initial analysis, the model proposed for the actuator has a similar configuration as the one presented in (Bueno and Leite Filho, 2003). However, further analysis of the step response indicates a slightly different third order linear model, given by the transfer function in Equation (1).

$$TF = \frac{305500}{s^3 + 202.1s^2 + 14520s + 326800} \quad (1)$$

Hence, the model becomes the one represented in Figure 1.

The HWIL simulation used for the limit cycle analysis consisted of a simplified dynamics model of the system followed by a PD controller (Bueno and Leite Filho, 2003), as seen in Figure 2. Since both the dead-zone and the backlash have known descriptive functions (Slotine and Li, 1991; Gelb and Vander Velde, 1968), the first harmonic analysis can be used to calculate the parameter values for those nonlinearities. Figure 3 shows the HWIL output for a given combination of K_p, K_d and μ_b.

Considering K_p=5.84, K_d=0.062 and μ_b=12.3 as the controller parameters, and assuming a time delay of T_d=0.0056, one finds f=1.1625e-04 for the backlash and δ=0.0092 for the dead-zone.

Those results correctly represent the limit cycle in both frequency and amplitude. However, as described by (Bueno and Leite Filho, 2003), this model has been unable to reproduce the shape of the signal encountered on the hardware-in-the-loop tests.

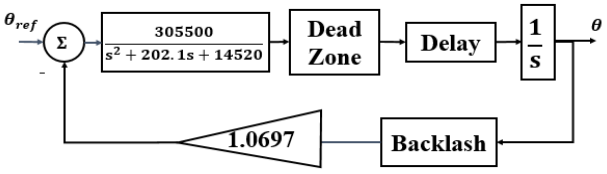


Figure 1. Initial configuration for actuator model.

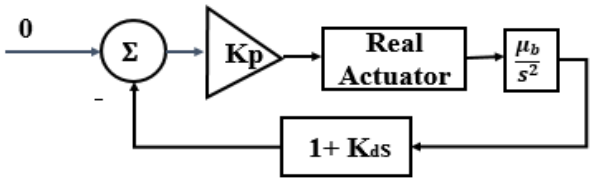


Figure 2. HWIL configuration for limit cycle analysis.

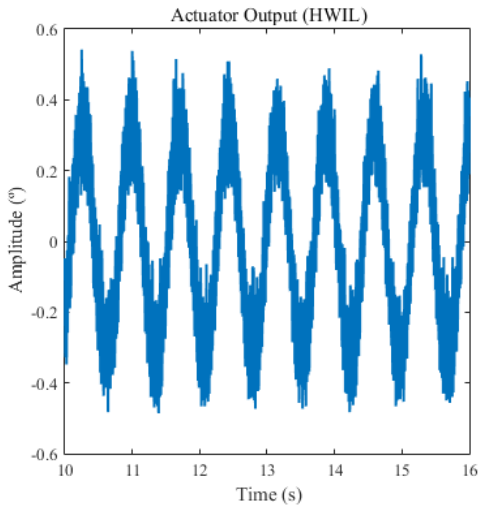


Figure 3. Actuator output (HWIL) for K_p=5.84, K_d=0.062 and μ_b=12.3.

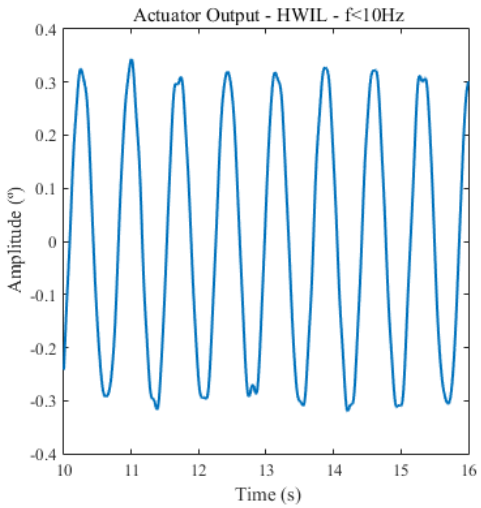


Figure 4. Actuator output (HWIL) with f<10Hz.

3 FFT Analysis

It is possible to reconstruct a signal using a finite Fourier series. This can be done by using the discrete Fourier transform (DFT) - an interpolating method capable of calculating the unknown coefficients for the series given a finite sample (Chu, 2008).

However, calculating the DFT directly is generally a procedure of order N^2 and it is not advisable (Chu, 2008). Instead, the most common approach is to use the Fast Fourier Transform (FFT), an algorithm capable of calculating the DFT of a sample of complex N data points with a speed proportional to $N\log_2 N$ (Cooley et al, 1967; Van Loan, 1992).

In order to obtain a more accurate look at the phenomenon studied, this work uses the Fast Fourier Transform (FFT) of the simulation output signal to better understand its properties. Once the frequency range of interest is identified, the Inverse Fast Fourier Transform (IFFT) can be used to reconstruct the signal without the influence of higher frequency noise.

The FFT analysis of the signal generated by the HWIL simulation has shown that frequencies above 10Hz could be ignored. An IFFT was then created so that the shape of the actuator output could be studied without the influence of external noise, as shown by Figure 4.

Figure 4 shows that the actual output presents periodic changes in its shape around the wave's antinodes, something that was not reproduced by the previous model. This indicates the existence of a relevant nonlinear phenomenon occurring when the actuator output changes direction of motion.

The physical model of the actuator guarantees the existence of an integral on the model, as shown on Figure 1 (Moreira and Leite Filho, 1988; Gibson, 1963). Therefore, it is reasonable to assume that this occurs when the derivative of the output crosses zero.

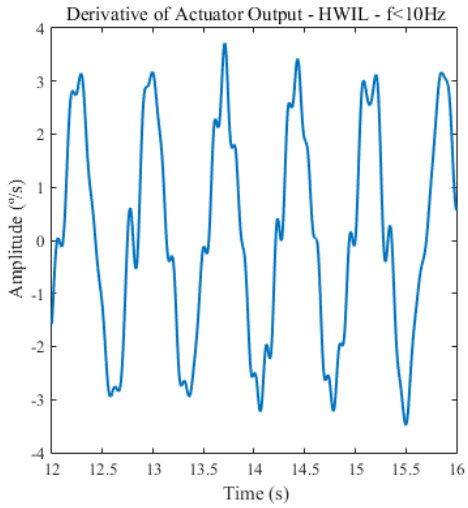


Figure 5. Reconstructed derivative.

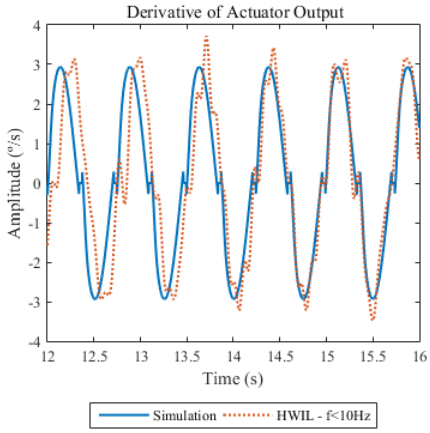


Figure 6. Derivative of actuator output including coulomb friction.

Based on the IFFT generated for frequencies smaller than 10Hz, as presented on Figure 4 and assuming that as the actual actuator output, it is possible to reconstruct the signal before the integral block, as shown on Figure

5. This strategy emphasizes the nonlinearities that one wants to model. If this signal can be reproduced by the inclusion of new nonlinearities, the new model will be able to recreate the real tests.

The spikes around zero on Figure 5 indicate the existence of an offset which sign opposes the direction of the derivative. This phenomenon can be reproduced with the inclusion of a new nonlinearity, modeled as a negative Coulomb friction, before the integral block. Figure 6 shows the effect of this element on the derivative signal simulated for a given set of parameters.

Since the simulated results are still not able to reproduce completely the HWIL signal, further analysis of the derivative is necessary. In order to better understand the relevant frequencies acting on the derivative signal, the Fast Fourier Transform can be used.

Figure 7 shows a graphic of the absolute value of the FFT result with respect to frequency. This graph shows that, while the simulated model presented proportional attenuation of the higher harmonics, the real actuator showed an increase in amplitude for frequencies between 9 Hz and 10Hz, especially around the seventh harmonic (9.57 Hz).

Therefore, it is important to be able to represent this phenomenon in order to reproduce the real results. The presence of higher harmonics seems to indicate that those were being stimulated somewhere on the actuator. As a way to recreate this on the model, a feedback loop is proposed.

The feedback loop must be able to affect only that specific frequency band, which must be amplified somewhere on the closed loop. Thus, a feedback loop with a bandpass filter is included on the model with an appropriate gain so that the results would match the HWIL tests.

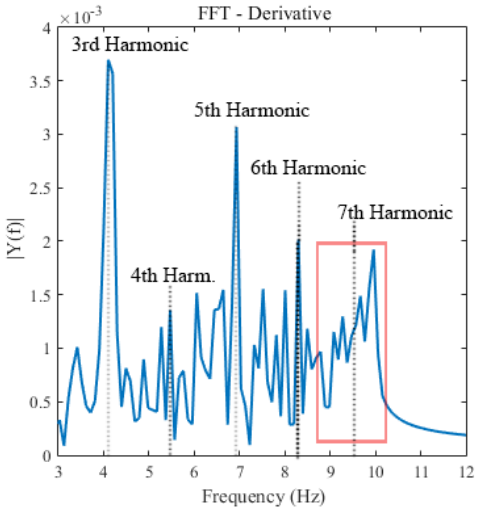


Figure 7. Absolute value of FFT – derivative of HWIL output (f<10Hz)

The presence of the feedback loop, however, influences the step response of the model, creating an oscillating signal that does not exist in reality. As a way to attenuate this, a saturation block is added so that the feedback loop will not falsely stimulate the system when given a nonzero input.

4 Model Structure

The final model proposed is shown on Figure 8. The filter bandpass used was a 4th order Butterworth design, with frequencies between 9 and 10 Hz.

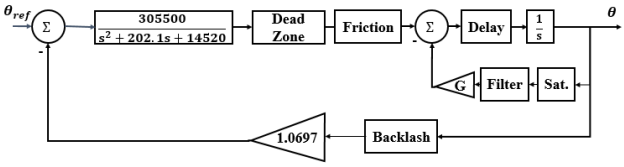


Figure 8. Final configuration for actuator model.

The presence of nonlinearities involving energy storage, such as friction, requires the use of a numerical approach in order to find the describing function (Duarte and Tenreiro Machado, 2006). Therefore, an analytical analysis no longer can be used to find the parameters that would recreate the limit cycle.

However, once a proper structure is found, different parameter values can be simulated until the response matches the HWIL results.

5 Simulation Results

For K_p=5.84, K_d=0.062 and μ_b=12.3, the actuator parameters were tuned so that f=1.1625e-04, δ=0.0083, T_d=0.0017, offset=-0.0054, G=48 and sat=0.015.

5.1 Limit Cycle Analysis

Initially, the model was validated by simulating the limit cycle under a PD controller in a similar configuration as described by Figure 2. The simulation results were compared to the HWIL results, analyzing both the output signal, its derivative and, finally, the absolute value of its FFT result with respect to frequency.

Figure 9 represents the derivative for both the HWIL signal and the simulation results, while Figure 10 illustrates the actuator output in both cases. Figure 11 and Figure 12 show, respectively, the absolute value Fast Fourier Transform of the derivative and the output for both the HWIL tests and the simulation results.

These results show that the new model is indeed able to recreate the limit cycle desired in frequency, amplitude and shape, showing an improvement when compared to the previous model.

However, there appears to be a periodic shift in phase that was not accounted for in this work.

Finally, since there might be more than one set of parameters that would reproduce the same limit cycle, further validation is required.

Thus, the final configuration is used in simulations with different values of K_p, K_d and μ_b, in order to verify if those are able to recreate the HWIL results. Figure 13 exemplifies the actuator output for one of those simulations.

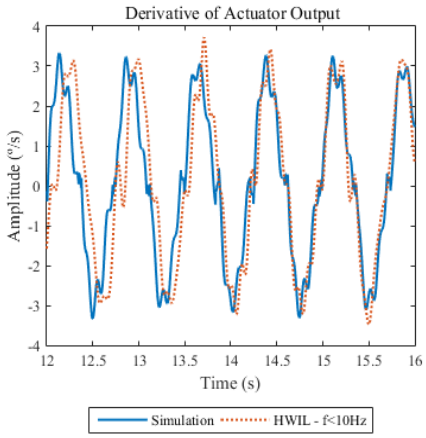


Figure 9. Derivative of actuator output for final configuration.

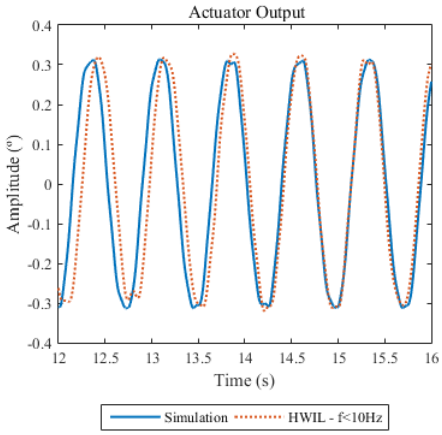


Figure 10. Actuator output for final configuration.

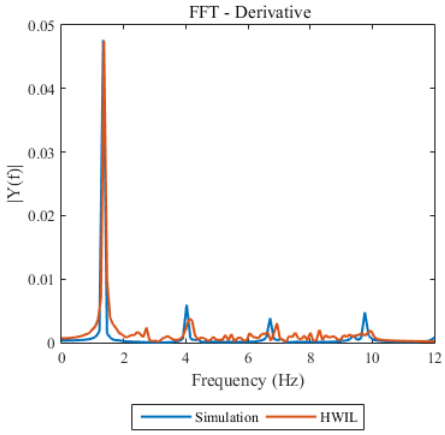


Figure 11. Absolute value of FFT of derivative of actuator output.

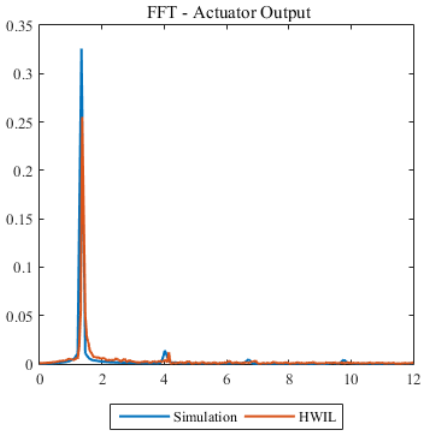


Figure 12. Absolute value of FFT of actuator output.

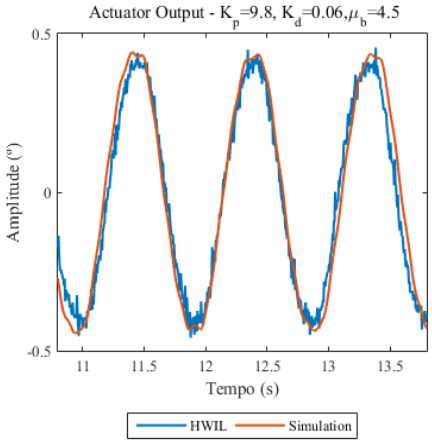


Figure 13. Actuator output for K_p=9.8, K_d=0.06 and μ_b=4.5.

5.2 Input Response

As a way to validate the model outside of the limit cycle conditions, different inputs were simulated and the outputs compared to results from tests on a real actuator.

Simulations were made for both square and sine wave inputs. The results are shown, respectively, on Figure 14 and Figure 15, respectively.

Since the delay block was assumed to be positioned before the integral block and, therefore, inside the closed loop, the values of T_d affected the shape of the output for a given square wave input. Thus, the acceptable values of transport delay are limited, which is why the final value used (T_d=0.0017) is smaller than the value considered for the initial model (T_d=0.0056).

As seen on Figure 15, this limiting factor has consequences on the output for the sinusoidal wave, where the actual actuator presents a higher delay than this model can reproduce.

A possible solution for this problem is to move the time delay block to after the feedback loop. This would solve the issue regarding the shape of the response to a square wave input and allow larger values for Td.

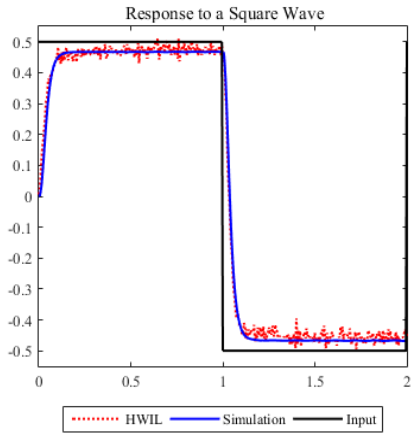


Figure 14. Actuator response for a square input.

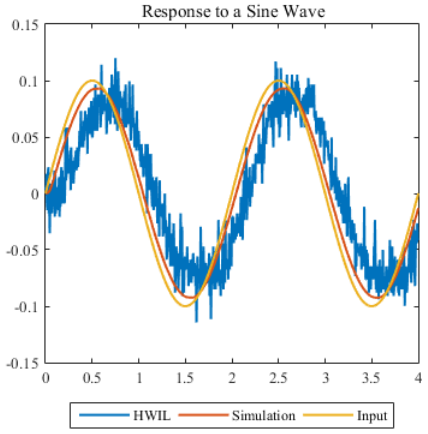


Figure 15. Actuator response for a sinusoidal input.

However, when this was implemented, no combination of parameters could be found where the results would be reproduced for all the available sets of K_d, K_p and μ_b. In most cases, when the results could be reproduced for a given set of controller parameters, the output for a different set would either generate wrong amplitudes or create system instability.

6 Conclusions

This paper presents a successful scheme of inferring a possible nonlinear configuration of a model based on the analysis of the FFT response of a reference signal and a complete simulation of the limit cycle. The validation of the parameters chosen for the actuator model is made by checking how the model's limit cycle responded to several HWIL parameters, as well as different inputs.

This study is based on data from a real actuator used for thrust vector control as part of the Brazilian Satellite Launcher (VLS) (Leite Filho, 1999; Leite Filho and Bueno, 2003), and the model created can be used to improve its control algorithms.

The new model was able to reproduce the HWIL results in both amplitude, frequency and shape, unlike the previous model proposed in (Bueno and Leite Filho, 2003). However, it is important to note that there seems to be a periodic phase shift along the output that was not completely reproduced by the simulation results. It is possible that this occurs because of a misplacement of the time delay block, but the model, as it stands, can be used for control systems simulations, per the original goal.

References

A. M. Bueno and W. C. Leite Filho. Parameter identification of actuator nonlinear model based on limit-cycle phenomenon. *Proc. 17th Int. Congr. Mechanical Engineering*. 2003.

Eleanor Chu. *Discrete and Continuous Fourier Transforms: Analysis, Applications and Fast Algorithms*. CRC Press. 2008.

James W. Cooley, Peter A. W. Lewis and Peter D. Welsh. Historical notes on the fast Fourier transform. *IEEE Transactions on Audio and Electroacoustics*, 15(2):76-79, 1967.

Fernando B. M. Duarte and J. A. Tenreiro Machado. Fractional dynamics in the describing function analysis of nonlinear friction. *Proc. 2nd IFAC Workshop Fractional Differentiation and its Applications. IFAC Proceedings Volumes*, 39(11):218-223, 2006.

Arthur Gelb and Wallace E. Vander Velde. *Multiple-Input Describing Functions and Nonlinear System Design*. McGraw-Hill. 1968.

J. E. Gibson. *Nonlinear Automatic Control.* McGraw-Hill, 1963.

W. C. Leite Filho. Control system of Brazilian launcher. *Proc. of 4th ESA Inter. Conf. on Guidance, Navigation and Control Systems*. pp. 401-405, 1999.

W. C. Leite Filho and A. M. Bueno. Analysis of limit-cycle phenomenon caused by actuator's non linearity. *Proc. of 12th IASTED Inter. Conf. on Applied Simulation & Modeling*. pp. 484-489, 2003.

F. J. O. Moreira and W. C. Leite Filho. Identificação de um servomecanismo de uma tubeira móvel. *Ann. VII Congresso Brasileiro de Automática*. pp. 383-387, 1988.

J. Slotine and W. Li. *Applied Nonlinear Control*. Prentice-Hall. 1991.

Charles Van Loan. *Computational Frameworks for the Fast Fourier Transform*. SIAM. 1992.

Mathematical Model of the Distribution of Laser Pulse Energy

Pavels Narica Artis Teilans Lyubomir Lazov Pavels Cacivkins Edmunds Teirumnieks
Faculty of Engineering, Rezekne Academy of Technologies, Latvia,
laser@rta.lv

Abstract

Method allows for modelling of the complex process of laser pulse energy distribution over flat work surface. The process of calculating the correct result does not use common lasing formulas but instead employs the mathematical model of matrix multiplication of three input matrices representing a pulse model, a line model, and a plane model. The pulse model represents the distribution of planar energy densities within the laser pulse. The line model represents the distribution of pulses within the line. The plane model represents the distribution of lines within the plane. Because mathematical model is implemented within a spreadsheet processor, its size can be adjusted as needed and it can be instantiated multiple times for simultaneous modelling of different input parameters.

Keywords: mathematical model, modelling method, laser pulse, energy distribution, planar energy density, matrix multiplication, laser marking, spreadsheet processor

1 Introduction

The main goal of this research is to simplify the process of understanding and visualizing the distribution of laser pulse planar energy densities over flat work surface.

Laser systems are widely used and provide some spectacular capabilities in many different fields. However lasers are complicated systems and operating a laser system requires a very good understanding of how lasing is actually done and how it can affect surroundings.

In scientific literature and on the internet there is a plenty of information about the laser systems. Such information contributes to better understanding of the lasing processes and includes some of the most widely used formulas and concepts, such as distance between two consecutive pulses, distance between two consecutive lines of pulses, individual laser pulse energy, and average lasing power (Bliedtner *et al*, 2013). However, the problem with such formulas and concepts is that they do not necessarily help one visualize the process being calculated. Another issue is that lasing processes are expensive.

Laser system usually provides for its operator a set of technical parameters which can be adjusted to obtain the necessary results. Nevertheless, these technical parameters may not provide a clear understanding of how they affect the lasing results. Even though laser system's operator might use provided technical parameters in formulas, the results of such formulas per se do not ensure being sufficiently useful.

In practice there are two main types of laser systems - pulsed lasers and continuous-wave (CW) lasers (Eichler, 1998). The former deliver energy to the work surface in discrete packets called pulses, while the latter emit photons continuously. Many of commonly used formulas are better suited for CW type laser systems, as these laser systems produce more predictable results. When such formulas are used in relation to pulsed laser systems, the results of formulas usually contain averaged values without local minimums and maximums.

Local minimums and maximums occur during both pulsed and CW type lasing and play an important role on produced results, as regions of work surface that are exposed to higher planar energy densities would behave differently than other remaining regions (Laakso *et al*, 2009; Antonczak *et al*, 2014). This is especially the case when producing colour laser marking on metals, as formed thin oxide films differ in regions exposed to higher planar energy densities compared to those exposed to lower planar energy densities, and thus have different appearance which has its contribution to overall perceived colour of marking (Ming *et al*, 2008; Veiko *et al*, 2014).

Pulsed laser systems emit pulses with some pulse repetition rate in the direction of scanning. Pulse repetition rate, scanning speed, the direction of scanning, pulse width, average lasing power, distance between two consecutive lines, and many other technical parameters are all set by laser operator. Because there are time periods between each two consecutive laser pulses, when no additional energy is delivered to the work surface, each two consecutive pulses may overlap in many different ways and thus distribute their energy over the work surface in many different forms. Each such distinct form of energy distribution in the end can affect work surface differently.

As already stated above, commonly used formulas do not allow for clear and instant understanding of how distribution of laser emitted energy over flat work surface would look like in practice and where its local maximums and minimums would be located. The best results common formulas can help to achieve, when given values of main technical parameters are known and laser construction specifics are taken into account, are to find distances between each two consecutive pulses and each two consecutive parallel lines of pulses as well as energy content of individual laser pulse and average lasing power that represents the rate of laser emitted energy delivery to the work surface.

When one knows the total amount of laser pulses delivered to a unit area of work surface as well as the energy content of each individual pulse, one may further calculate average planar energy density for that unit area. However, just by using common formulas it would not be easy to obtain information about the actual distribution of different planar energy densities, as these can differ based on their position on the work surface due to the overlapping effects between each two consecutive pulses and each two consecutive parallel lines of pulses.

The present mathematical model allows for modelling of laser pulse energy distribution over flat work surface. In particular, it is a novel method that deploys built-in spreadsheet processor's matrix multiplication function in order to automatically generate informative numeric data in form of a two-dimensional histogram that can be further used for visualizing the actual distribution of laser pulse planar energy densities over flat work surface.

2 Materials and methods

Experiments were carried out using PowerLine F-20 Varia series pulsed fiber laser system produced by Rofin-Sinar Laser GmbH. It emits photons of wavelength equal to 1064 nm, has maximum average lasing power of 20 W, pulse repetition rate of 2-1000 kHz, adjustable pulse width of 4-200 ns.

The colour palette shown in Figure 1 contains the yellow colour which is also called sample colour 4 throughout the text. The image in Figure 2 was taken using optical microscope Meiji Techno MT and represents the sample colour 4. The stainless steel sample used for colour laser marking was 4301 18-9E 2R.

The sample colour 4 described is Figure 1, Figure 2, Figure 3, and Figure 4 has the following technical laser parameters associated with it: pumping power of 25% (equivalent of 2 W average power, given specified pulse repetition rate and pulse width), pulse repetition rate of 200 kHz, scanning speed of 200 mm/s, pulse width of 4 ns, and distance between two lines of 5 μm.

Figure 1. Produced sample marking colours on stainless steel workpiece.

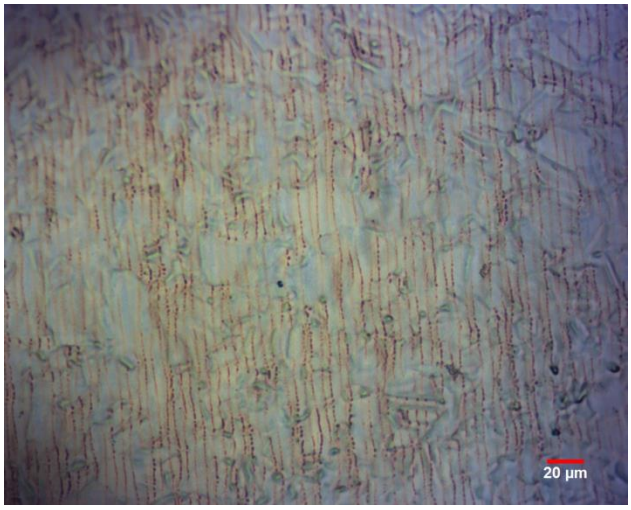


Figure 2. Sample marking colour 4 under optical microscope.

The method of modelling the distribution of laser pulse planar energy densities over flat surface allows for analysis of lasing processes by providing a mathematical model which consists of six related matrices. No common lasing formulas are used by the mathematical model, though any model's user can always derive necessary lasing formulas from the model by analyzing its state. The present method can provide different kinds of relevant information about laser pulse energy distribution over some flat surface, such as linear energy densities, linear pulse densities, distance between two consecutive pulses or lines of pulses, individual laser pulse energy.

The aim of the method is to make the most prominent feature of laser systems easier to understand, define, quantify, visualize, teach, and simulate by referencing it to existing and usually commonly accepted knowledge and formulas. The method offers a new way of looking at what actually happens with laser pulses, as they are being accumulated on some work surface.

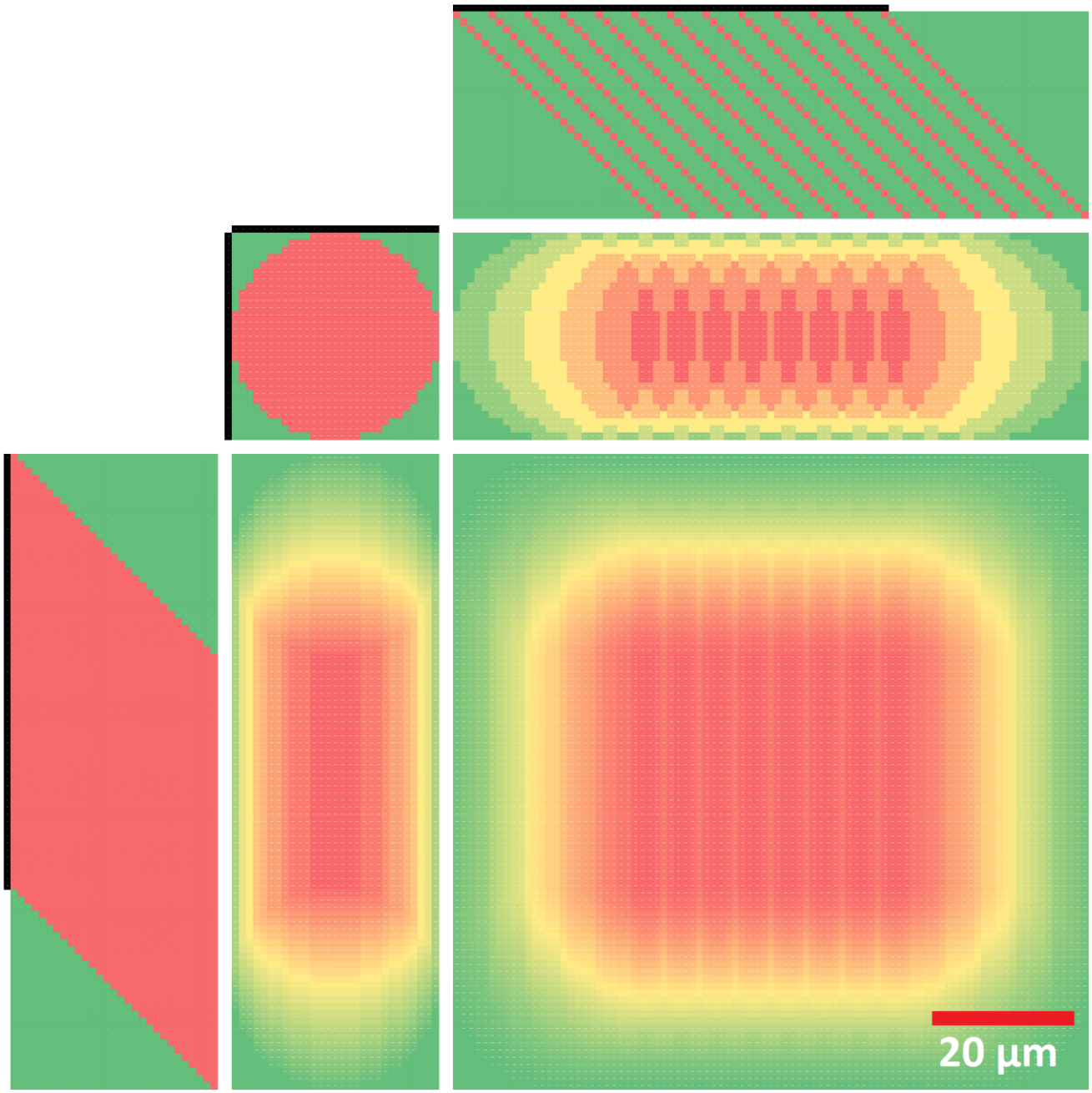


Figure 3. Produced visualization of the distribution of laser pulse planar energy densities for sample colour 4.

The method allows users to simulate actual distribution of planar energy densities over flat work surface. This is accomplished by interacting with model's input area consisting of numeric square matrix "P" (Figure 5), dedicated for defining distribution of planar energy densities within the pulse itself, as well as two perpendicular numeric vectors "y" and "x", the former allowing to define distribution of pulses within a line and the latter allowing to define the distribution of these lines within a work surface plane. Numeric input of column vector "y" is automatically copied into each remaining column vector of matrix "Y" and shifted along the length dimension of the matrix such that diagonal lines of identical numbers are formed. Numeric input of row vector "x" is automatically copied into each remaining row vector of matrix "X" (Figure 6) and shifted along the width dimension of the matrix such that diagonal lines of identical numbers are formed.

The most important aspect of the method is that it correctly simulates the actual distribution of planar energy densities over flat work surface from the aspect of both – physics and mathematics. By multiplying three matrices "Y", "P", and "X" the mathematical model is able to calculate positions of delivery of every laser pulse relative to the flat work surface. The distribution of laser pulses is immediately displayed back in form of three output matrices "YP", "PX", and "YPX". The numeric information of matrix "YPX" can easily be visualized as a three-dimensional surface

chart to display the distribution of laser pulse planar energy densities (Figure 4).

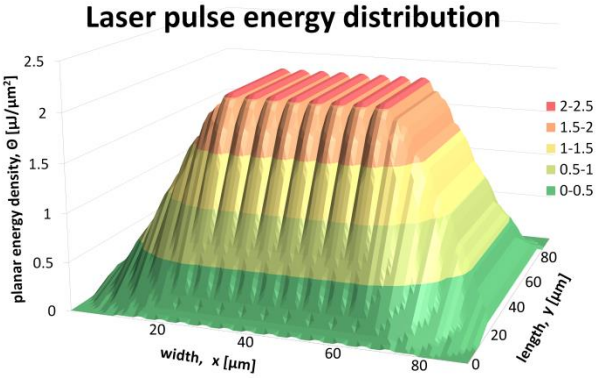


Figure 4. Three-dimensional surface chart of the distribution of laser pulse planar energy densities for sample colour 4.

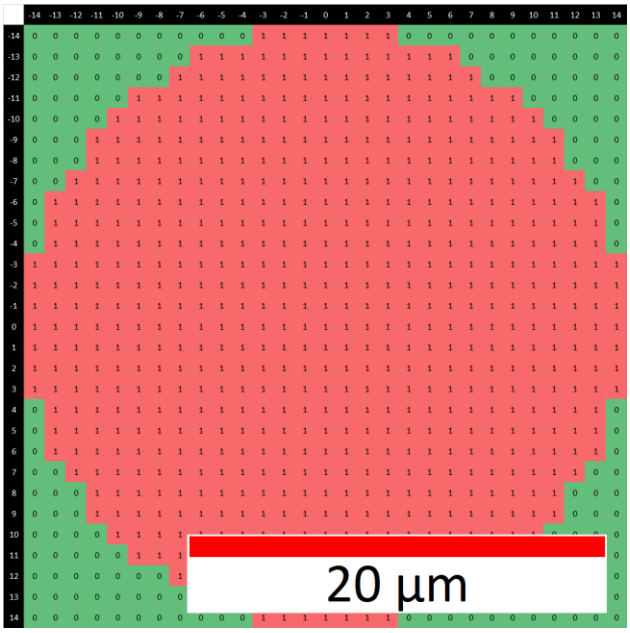


Figure 5. Example of pulse model "P".

3 Results and discussion

The method of modelling the distribution of laser pulse planar energy densities by the use of matrix multiplication is implemented by the mathematical model comprised of six related matrices - three input matrices and three output matrices (Figure 7). The input matrices accept user provided data in form of numbers. The output matrices update their numeric states every time user makes changes to the input matrices.

The method always produces correct results because of its simple underlying logic that photons are both carriers and units of energy and that they are additive. Thus laser pulse is a representation of some quantity of photons, and laser pulse total energy is always proportional to the amount of photons it consists of due to energy being an extensive physical property. The mathematical model stores one such laser pulse as a model itself in a form of a square matrix "P". It consists of numeric data that represent arbitrary amounts of photons and their relative positions within that laser pulse's planar surface projection. The mathematical model then uses this user defined or default laser pulse model to copy and to distribute it over the flat work surface.

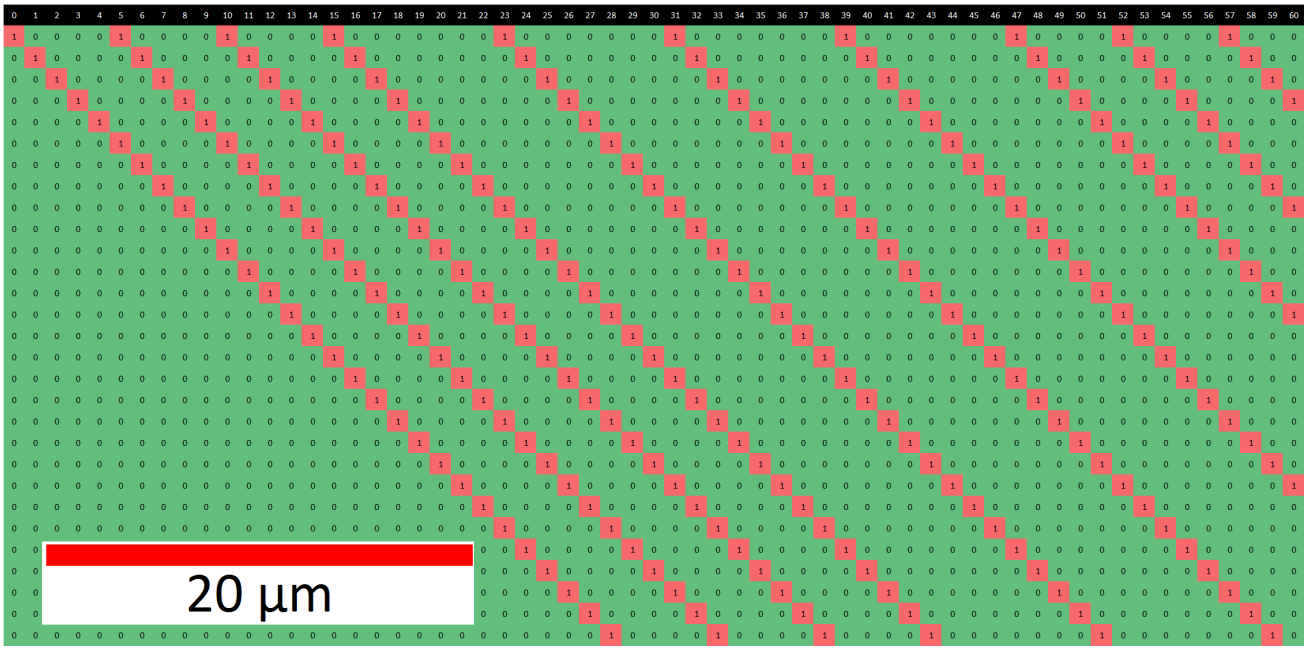


Figure 6. Example of line model "X".

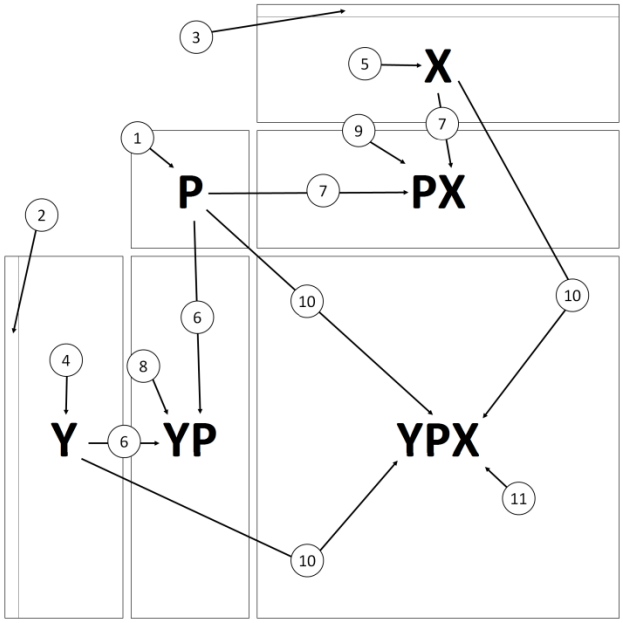


Figure 7. Schematic view of mathematical model implemented by the present method.

Besides input matrix "P" there are two perpendicular input matrices "Y" and "X" representing line model and plane model respectively. Both input matrices "Y" and "X" serve identical functions of storing positions of where laser pulses or lines of pulses are to be delivered in relation to the work surface. Input matrix "Y" is associated with flat work surface's length dimension, while input matrix "X" is associated with its width dimension. When viewed, the input matrix "Y" appears to have vertical rectangle shape with its width equal to the side length of square matrix "P", while the input matrix "X" appears to have horizontal rectangle shape with its length equal to the side width of square matrix "P". The work surface is itself represented by an output matrix "YPX" with its length equal to the length of matrix "Y" and its width equal to the width of matrix "X".

The mathematical model calculates three different matrix products and stores them in the output matrices "YP", "PX", and "YPX" to display the distribution of laser pulse planar energy densities from three different points of view: as two perpendicular to one another laser pulse lines "YP" and "PX", and as a matrix representation of work surface "YPX".

Both input matrices "Y" and "X" are based on the idea of the identity matrix which is a special case of matrix in mathematics that does not change other matrices it is being multiplied with. Input matrices "Y" and "X" are not themselves identity matrices, yet they share some very important similarities - only diagonal lines of numbers are stored in them and each distinct diagonal line consists of identical numbers specified by mathematical model's user within the first column vector "y" of input matrix "Y" and the first row vector "x" of input matrix "X".

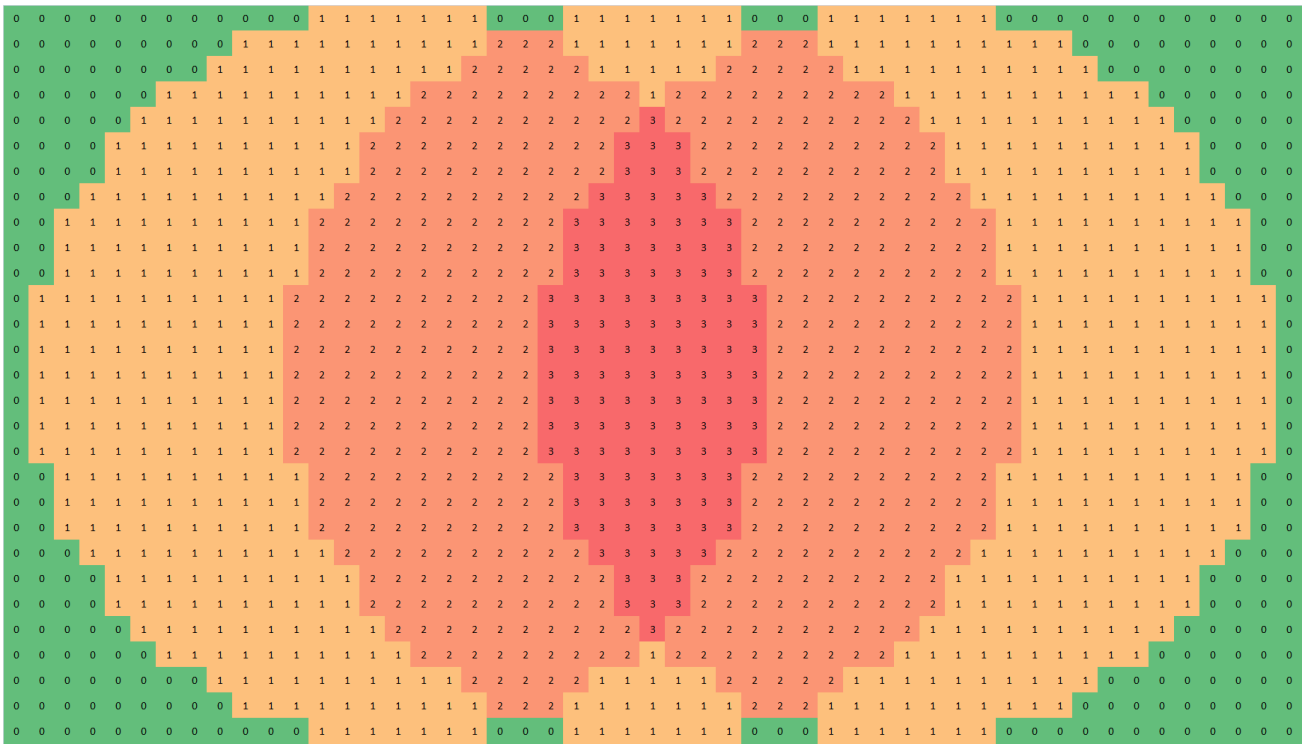


Figure 8. Example sum of three partly overlapping laser pulses.

Such energy content addition is in accordance with the laws of physics and is provided by mentioned procedure of multiplication of input matrices. When three or more pulses are to be somehow overlapped on work surface, then matrix multiplication sums energy contents of all pulse overlapped areas based on how many pulses share that area simultaneously.

All six matrices that constitute the mathematical model share one very important common property - each matrix consists of spreadsheet processor's cells. All the cells within the spreadsheet processor's worksheet which contains full instance of the mathematical model have square shape and the same area. This property allows one to refer to each cell as a unit area used throughout current instance of the mathematical model. Each such unit area has four sides of equal unit length.

The ability to refer to mathematical model cells by their common unit area and unit length allows for the model user to measure processes and states within input and output matrices using any preferred arbitrary base unit of length. Because all data within model matrices are of a numeric type, user can select multiple adjacent cells within any of six matrices to see the sum, average, minimum, maximum, and other statistical values representing user selection with the help of spreadsheet processor's standard features.

The sum of numeric content of all cells within input matrix "P" represents the total amount of user-defined or the model's default laser pulse energy which as well can be expressed by user in any preferred base units. By using standard spreadsheet processor function for counting all non-zero cells within matrix "P" one can quickly find the total area of laser pulse planar projection on work surface. By selecting any individual cell of output matrix "YPX" one can view the total amount of energy accumulated by the corresponding unit area of work surface.

One can assume that there will always be potential for finding more new ways of extracting additional information about the distribution of laser pulse planar energy densities. The present method allows its users not only to model the distribution of laser pulse planar energy densities for both laser system types - pulsed and CW - but also to combine several different instances of mathematical model by summing all the necessary output matrices "YPX" throughout these open instances and outputting the sum into a new spreadsheet processor's worksheet. Such technique would allow the model user to render each consecutive step of laser pulse delivery to the work surface so that a complex animation can be produced. One can even model processes that are not so easily achievable with common laser systems and their technical parameters.

The user of the mathematical model would often view output matrices zoomed out because of their size thus all numeric data in all the matrices are colour-scaled using spreadsheet processor's standard conditional formatting feature so that numbers are visually represented as colours depending on their relative magnitudes within their corresponding matrices.

4 Conclusions

The user's overall understanding of how the present method's mathematical model of matrix multiplication works can help the user extract even more useful information out of it. Therefore method can be used for interactive teaching purposes or to assist advanced users. Finally, the method can help its users better to interpret and to test common lasing formulas as well as to produce new ones, and no similar modelling method, which allows that, yet exists.

During testing of the mathematical model it was found that the better the distribution of planar energy densities within pulse model is defined the more accurate results are produced on the output. The same applies to the resolution, as the smaller unit areas produce better results.

Before the final model was developed, its previous version was implemented using HTML5 Canvas and JavaScript technologies, and pulse overlapping was achieved by visually combining semi-transparent circles, though output results did not provide any numeric data. Therefore final model is based on numeric data, and visualizations are model's by-products. Nevertheless these visualizations can help one spot many important patterns, such as recurring rectangular patterns of length equal to distance between pulses and width equal to distance between lines within output matrix of flat work surface.

Acknowledgements

This research was supported by laser system equipment from Rofin-Sinar Laser GmbH.

References

A. J. Antonczak, B. Stepak, P. E. Kozioł, and K. M. Abramski. The influence of process parameters on the laser-induced coloring of titanium. *Applied Physics A*, 115(3):1003-1013, 2014. doi: 10.1007/s00339-013-7932-8

J. Bliedtner, H. Muller, and A. Barz. Lasermaterialbearbeitung. *Fachbuchverlag Leipzig im Carl Hanser Verlag*, 2013. doi: 10.3139/9783446429291

J. Eichler and H. J. Eichler. Laser. *Springer-Verlag Berlin*, 1998. doi: 10.1007/978-3-662-08247-8

P. Laakso, S. Ruotsalainen, H. Leinonen, A. Helle, R. Penttilä, A. Lehmuskero, and J. Hiltunen. *Direct color marking of metals with fiber lasers*. VTT, Lappeenranta, Finland, 2009. Research Report VTT-R-02403-09

L. Ming, A. Tse, and T. Hoult. Colour marking of metals with fibre lasers. *In Proceedings of the 3rd Pacific International Conference on Application of Lasers and Optics*, 2008.

V. Veiko, G. Odintsova, E. Ageev, Y. Karlagina, A. Loginov, A. Skuratova, and E. Gorbunova. Controlled oxide films formation by nanosecond laser pulses for color marking. *Optics Express*, 22(20):24342-24347, 2014. doi: 10.1364/oe.22.024342

Mathematical Model of Forecasting Laser Marking Experiment Results

Pavels Narica Artis Teilans Lyubomir Lazov Pavels Cacivkins Edmunds Teirumnieks
Faculty of Engineering, Rezekne Academy of Technologies, Latvia,
laser@rta.lv

Abstract

Method allows for modelling of the anticipatory results of colour laser marking experiments. The process of calculating expected results takes into consideration the construction specifics of laser system being used and displays the results in compact form of a set of parameter matrices that have their values conditionally formatted as colour maps for easy identification of complex patterns. The complete set of all the related parameter matrices, both technical and derived, as well as the specific relations between them form the mathematical model of forecasting laser marking experiment results. Because the mathematical model is implemented within spreadsheet processor, it can be instantiated multiple times for any number of experiments.

Keywords: mathematical model, modelling method, laser system construction specifics, visualization of parameters, matrices, experiments, colour laser marking

1 Introduction

The main goal of this research is to simplify the process of forecasting and visualizing the distribution of values of experimental laser parameters.

Laser systems are widely used and provide some spectacular capabilities in many different fields. However lasers are complicated systems and operating a laser system requires a very good understanding of how lasing is actually done and how it can affect surroundings.

One case of common use of laser systems is colour laser marking on metal. Pulsed fiber lasers are usually used for producing markings of different colours on metal workpiece. Such laser systems are rather small in size, very precise, relatively inexpensive, and provide easy to use computer software for plotting laser marking elements. When plotting different elements in marking software, different technical laser parameters can be assigned to them in order to produce different marked colours (Laasko *et al*, 2008; Antonczak *et al*, 2014; Veiko *et al*, 2016; Qia *et al*, 2003; Antonczak *et al*, 2013; Amara *et al*, 2015; Lehmuskero *et al*, 2010; Veiko *et al*, 2014).

In many cases while operator is testing the laser system for its marking capabilities he/she marks some matrices of different combinations of laser's technical or other physical parameters. While this is logical approach, in most such cases it has serious issues, because the construction of laser system may behave differently under different set of parameters and thus produce unexpected or difficult to interpret marking results.

Because metal workpieces used for laser marking tests usually have rectangular flat two-dimensional surface (Figure 1), plotting on them rectangular matrices of rectangular elements is very efficient. This way an operator can assign for all the rows of such matrix some distinct technical or other test parameter and then assign all columns some other distinct test parameter while at the same time varying linear or logarithmic values of both assigned test parameters for each row and each column.

Figure 1. Experimentally produced sample marking colours on stainless steel workpiece.

When both dynamic test parameters are assigned to rows and columns of the matrix, all remaining parameters for all matrix elements are set to have some constant value. Such plotted matrix is then marked and can be further analyzed. This way one should be able to see how varying values of only two test parameters affect the produced marking colours. Though, as it was mentioned above, in some cases the construction of laser system itself can indirectly affect some additional parameters thus making analysis of marking colours against selected dynamic test parameters of marked matrix much more difficult and unintuitive.

When output marking samples are analyzed, the corresponding output colours are usually mapped against technical and other physical parameters set dynamically before marking test as well as other constant technical laser parameters and parameters related to the nature of marking process itself. This approach helps one better see which parameters cause marking colours to change. Yet such mappings usually are in form of linear flat tables requiring lots of rows for each experimentally marked colour and columns for listing all parameters to display all the necessary data. Within such table it may become very hard for one to efficiently identify important parameters that actually affect the formation of marked colour.

Finally, it is common practice to produce marking experiment first and then to analyze resulting colours against parameters afterwards. Such practice can often result in regions within marked test matrix that do not have many distinct and quality colours in them.

The present method allows for modelling colour laser marking experiments while considering the construction specifics of a laser system. In particular, it is a novel method that deploys built-in spreadsheet processor's features in order to automatically generate informative numeric data matrices and their visualizations by means of conditional formatting (Figure 2, Figure 3).

Planar energy density θ [J/mm^-2]									
3.1	4	5	6.3	10	13	20	25	50	100
2.5	3.2	4	5	8	10	16	20	40	80
1.6	2	2.5	3.1	5	6.3	10	13	25	50
1.3	1.6	2	2.5	4	5	8	10	20	40
0.8	1	1.3	1.6	2.5	3.1	5	6.3	13	25
0.6	0.8	1	1.3	2	2.5	4	5	10	20
0.5	0.6	0.8	1	1.6	2	3.2	4	8	16
0.4	0.5	0.6	0.8	1.3	1.6	2.5	3.1	6.3	13
0.3	0.4	0.5	0.6	1	1.3	2	2.5	5	10
0.3	0.3	0.4	0.5	0.8	1	1.6	2	4	8

Figure 2. Distribution of planar energy densities for sample.

2 Materials and methods

Experiments were carried out using PowerLine F-20 Varia series pulsed fiber laser system produced by Rofin-Sinar Laser GmbH. It emits photons of wavelength equal to 1064 nm, has maximum average lasing power of 20 W, pulse repetition rate of 2-1000 kHz, adjustable pulse width of 4-200 ns. The stainless steel sample used for colour laser marking was 4301 18-9E 2R.

Table 1 provides information used by current implementation of the mathematical model. It uses such common technical parameters as *f* for pulse repetition rate, *v* for scanning speed, *P* for average lasing power, *Δx* for line step (distance between two lines), *x* and *y* for width and length of individual rectangular marking element (Bliedtner *et al*, 2013; Eichler, 1998).

Table 1. Formulas used within proposed mathematical model as an example.

Description	*Formula*	*Base unit*
Pulse overlap (linear pulse density)	$N_y = \frac{f}{v}$	$[mm^{-1}]$
Line overlap (linear line density)	$N_x = \frac{1}{\Delta x}$	$[mm^{-1}]$
Pulse energy	$E_P = \frac{P}{f}$	$[\mu J]$
Planar pulse density	$N = N_y \cdot N_x$	$[mm^{-2}]$
Planar energy denstiy	$\Theta = E_P \cdot N_y \cdot N_x$	$[J/mm^{-2}]$
Total energy delivered to element	$E = \Theta \cdot y \cdot x$	$[J]$
Total marking time of element	$t = \frac{N_x \cdot x \cdot y}{v}$	$[s]$

The method of modelling colour laser marking experiments takes into consideration the construction specifics of laser system. The method then displays anticipated results in compact form using equally sized matrices that have conditional formatting applied to their values. Present method allows for quick visual relation of parameters and their values as well as identification of patterns associated with the experiment (Figure 3).

The aim of the method is to make analysis of the colour laser marking experiments easier to understand, define, quantify, visualize, teach, and simulate by referencing it to existing and usually commonly accepted knowledge and formulas. The method offers a new way of looking at what actually happens to different experimental parameters before the actual marking of experiments. Therefore, as one identifies the set of the most important experimental parameters affecting the colour laser marking process, one can then better optimize the colour laser marking process itself.

The method implements one primary mathematical model of forecasting results for colour laser marking experiments that incorporates three secondary interconnected mathematical models (Figure 4). The primary mathematical model is a set of related matrices (Figure 5) with relations being defined by lasing formulas shown in Table I. The present method allows for addition of new as well as removal of unnecessary experimental parameter matrices, both technical and derived. The present method allows for very fast and flexible input of technical parameters and their values for the experiment as well as immediate output of the expected results.

Figure 3. Overview of proposed mathematical model for experimentally produced sample (with constant pulse width of 4 ns).

Once all three secondary mathematical models are defined, the primary mathematical model is set and can be instantiated for any different experiment. Thus, before one proceeds with actual marking of test matrices, it is possible to efficiently plan the experiments and see expected distribution of values of output parameters in advance.

3 Results and discussion

The method maps and visualizes parameters of colour laser marking experiments. The method displays all the necessary experimental parameters, both technical and derived, that are associated with marking experiment of a test matrix in a compact form of a set of equally sized matrices where each such matrix is assigned its own experimental parameter and stores that parameter's values positionally related to elements of test matrix to be marked as well as the values of every other experimental parameter matrix. Thus the number of rows and columns of each such experimental parameter matrix is identical to those of test matrix itself.

The method of mapping values of necessary experimental parameters within their own matrices to elements of the test matrix lets one immediately start seeing experimental values in context of two dimensions corresponding to those of test matrix, also in the context of values belonging to all other experimental parameter matrices, and, more importantly, in the context of marking colours to be produced. Thus one might think of the set of all experimental parameter matrices including the test matrix to be marked itself as a three-dimensional stack of matrices where each matrix cell that corresponds to specific row and specific column is directly associated with every other such cell up and down the stack.

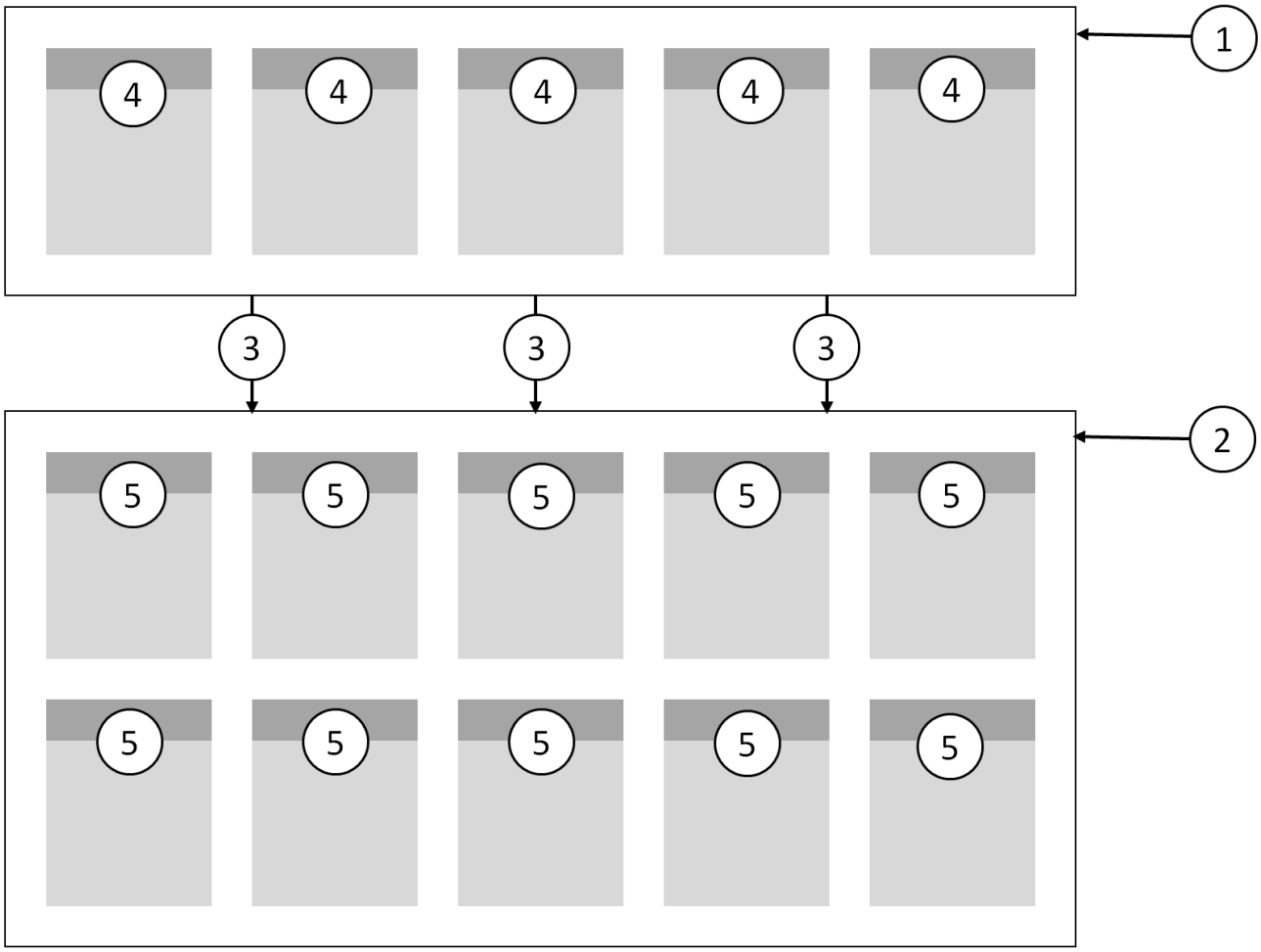


Figure 4. Schematic view of primary mathematical model implemented by the present method: (1) secondary mathematical model of input of experimental data; (2) secondary mathematical model of output of experimental data; (3) secondary mathematical model of laser system's construction specifics; (4) instances of laser system's technical parameter matrices; (5) instances of derived parameter matrices.

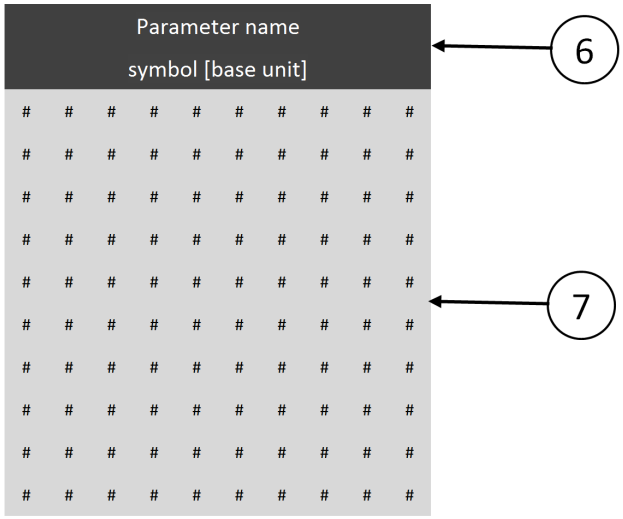


Figure 5. Schematic view of parameter's matrix: (6) header associated with parameter's matrix containing information about the parameter in question, its commonly accepted physical symbol and base unit; (7) two-dimensional array of distribution of values of the parameter.

By applying spreadsheet processor conditional formatting feature to the values of each experimental parameter matrix, one can see the planar distribution of relative magnitudes of these values coded with a colour scale. Thus each experimental parameter matrix becomes a colour map and can be analyzed when spreadsheet processor worksheet is zoomed out. Each such colour map helps efficiently spot contrasting values within all matrices just by quickly looking through them.

Interestingly enough, distinct patterns may emerge within some of such experimental parameter colour maps. Such patterns may even sometimes resemble the pattern seen within the marked test matrix itself. Though, more often similar patterns can be observed within the set of experimental parameter matrices. Thus if any two such patterns look similar or even identical one might assume these are somehow proportional to each other.

The method implements the primary mathematical model of forecasting results for colour laser marking experiments which in turn comprises three secondary

interconnected mathematical models. The secondary mathematical model of laser system construction specifics allows user to define any conditions where laser system may behave differently by setting specific relations between the set of technical parameter matrices and the set of derived parameter matrices. The set of all laser system's technical parameter matrices represents the secondary mathematical model for input of experimental data. The set of all derived parameter matrices forms the secondary mathematical model for output of experimental data.

Thus the method allows for modelling distribution of the values within experimental laser parameter matrices for any particular colour laser marking experiment. Both secondary mathematical models for input of experimental data and for output of experimental data operate with numeric data.

4 Conclusions

The user's overall understanding of how the present method's primary mathematical model of forecasting laser marking experiment results works can help the user extract even more useful information out of it. Therefore the method can be used for interactive teaching purposes or to assist advanced users. Finally, the method can help its users better interpret and test common lasing formulas as well as produce new ones, and no similar modelling method, which allows that, yet exists.

There is always a problem of experiment repeatability when testing produced marking colours for technical parameters mentioned in research papers of other authors. The proposed mathematical model should be able to help authors to present their experimental parameters and their values as well as produced marking colours in a very compact form that is both easy to understand and easy to implement.

During testing of the mathematical model it was found that different formulas which define relations between the technical parameter set and the derived parameter set are easy to view, edit and copy. The method can also be useful for researchers testing different theories about colour laser marking, since adding new or eliminating unnecessary parameter matrices is easy and different emerging colour patterns help quickly spot parameters that matter in the context of colour laser marking experiments.

Acknowledgements

This research was supported by laser system equipment from Rofin-Sinar Laser GmbH.

References

E. H. Amara, F. Haïd, and A. Noukaz. Experimental investigations on fiber laser color marking of steels. *Applied Surface Science*, 351:1-12, 2015. doi: 10.1016/j.apsusc.2015.05.095

A. J. Antonczak, D. Kocoń, M. Nowak, P. Kozioł, and K. M. Abramski. Laser-induced colour marking—Sensitivity scaling for a stainless steel. *Applied Surface Science*, 264:229-236, 2013. doi: 10.1016/j.apsusc.2012.09.178

A. J. Antonczak, B. Stepak, P. E. Kozioł, and K. M. Abramski. The influence of process parameters on the laser-induced coloring of titanium. *Applied Physics A*, 115(3):1003-1013, 2014. doi: 10.1007/s00339-013-7932-8

J. Bliedtner, H. Muller, and A. Barz. Lasermaterialbearbeitung. *Fachbuchverlag Leipzig im Carl Hanser Verlag*, 2013. doi: 10.3139/9783446429291

J. Eichler and H. J. Eichler. Laser. *Springer-Verlag Berlin*, 1998. doi: 10.1007/978-3-662-08247-8

P. Laakso, H. Pantsar, and V. Mehtälä. *Marking decorative features to stainless steel with fiber laser*. IMD/ALAC, 2008.

A. Lehmuskero, V. Kontturi, J. Hiltunen, and M. Kuittinen. Modeling of laser-colored stainless steel surfaces by color pixels. *Applied Physics B*, 98(2-3):497-500, 2010. doi: 10.1007/s00340-009-3734-2

J. Qi, K. L. Wang, and Y. M. Zhu. A study on the laser marking process of stainless steel. *Journal of Materials Processing Technology*, 139(1-3):273-276, 2003. doi: 10.1016/S0924-0136(03)00234-6

V. Veiko, G. Odintsova, E. Ageev, Y. Karlagina, A. Loginov, A. Skuratova, and E. Gorbunova. Controlled oxide films formation by nanosecond laser pulses for color marking. *Optics Express*, 22(20):24342-24347, 2014. doi: 10.1364/oe.22.024342

V. Veiko, G. Odintsova, E. Gorbunova, E. Ageev, A. Shimko, Y. Karlagina, and Y. Andreeva. Development of complete color palette based on spectrophotometric measurements of steel oxidation results for enhancement of color laser marking technology. *Materials & Design*, 89:684-688, 2016. doi: 10.1016/j.matdes.2015.10.030

Classification of OpenCL Kernels for accelerating Java Multi-agent Simulation

Pitipat Penbharkkul and Worawan Marurngsith
Department of Computer Science, Thammasat University, Pathum Thani, Thailand,

Abstract

Java-based multi-agent simulation (MAS) can be offloaded to graphical processing units (GPU) and other OpenCL accelerators to achieve many hundred-fold speedups. However, the performance gain from the accelerated code depends strongly on whether the computation (kernels) have been scheduled to the appropriate devices. Thus, accelerating Java MAS may not lead to a sustainable speedup. This paper proposes a method for a kernel classifier to specify suitable devices to execute OpenCL kernels. The classifier can identify suitable OpenCL devices for kernels based on the static and dynamic characteristics of the code of the kernels. Kernels are grouped by their suitability for particular devices using the multiclass support virtual machine technique. After that, kernels are scheduled to an appropriate task queue. Kernel scheduling based on the proposed technique is compared against the firstcome-first-serve (FCFS) technique and against oracle scheduling when handling eight kernels. Our results show that, using the proposed method, all kernels finished execution 45 percent sooner than using the FCFS technique. However, the overall execution time was 22.5 percent longer than with oracle scheduling. Our results seem to confirm that kernel classification techniques might contribute towards sustainable high performance in accelerated Java-based MAS models.

Keywords: GPGPU, OpenCL, multi-agent simulation, performance, acceleration, SVM, MASON

1 Introduction

Java is a powerful platform for developing multi-agent simulation (MAS) models due to its portability and the huge amount of functionality available in development kits. However, limitations in performance and scalability make Java-based MAS a target for performance acceleration on heterogeneous platforms *e.g.*, multicore CPUs, graphical processing units (GPUs), accelerated processing units (APUs), or coprocessors such as the Intel Xeon Phi (Aaby et al., 2010; Hayashi et al., 2013; Ho et al., 2015; Li et al., 2016). Offloading a cellular automata simulation, *e.g.* the Conway's Game of Life, to a cluster of GPUs could gain a 100x speedup if proper latency hiding techniques are used (Aaby et al., 2010). Well optimised MAS models based on the MASON library, a legacy Java-based MAS framework, could be accelerated from 100x up to 468x (Ho et al., 2015). Recent work proposing AgentPool and agent management techniques has shown that accelerated models can outperform CPU and GPU implementations based on the MASON and FLAME simulation frameworks (Li et al., 2016).

The OpenCL language layer (Group, 2013) has been widely used to exploit data parallelism on heterogeneous systems because of its portability and acceptable performance (Sachetto Oliveira et al., 2012). Research has shown that a single version of OpenCL code can be executed on three different platforms with less than a 12% performance loss, providing that parameters are well chosen (Dolbeau et al., 2013). Programmers can offload data-parallel fragments of Java code to heterogeneous platforms supporting standard OpenCL in two ways: by using an auto-parallelisation tool (AMD Developer Central, 2011; Hayashi et al., 2013) or by manually specifying fragments of parallelisable code using Java to OpenCL bindings such as JOCL (JOCL, 2011).

However, a well-known limitation of OpenCL is that the performance gain from accelerated code depends strongly on scheduling computation (kernels) to appropriate devices. This limitation also applies to accelerated Java MAS code. Hence, its speedup can be a hundredfold or zero. Several ways to predict performance of OpenCL kernels for different devices have been mentioned in three extensive surveys (Mokhtari and Stumm, 2014; Rossbach et al., 2013; Yan et al., 2009). Kernel profiling is a key technique used to get information about kernels to be classified, *e.g.* retrieving from history with profile data (Sato et al., 2011) or developing a framework for profiling a shared library (Matoga et al., 2013). A compiler framework to collect and classify kernels suitable for different devices was proposed in (Lopez-Novoa et al., 2015; Wen et al., 2014). These reports have confirmed that by using classification techniques, a kernel speedup on different OpenCL device can be predicted in advance at up to 87 percent accuracy (Wen et al., 2014)

Table 1. Parallelisation Techniques for ABS.

ABS Class	Parallelisation Techniques
Homogeneous	Agent's computation is implemented as kernel (Ho et al., 2015; Li et al., 2016).
Heterogeneous	Agent's computation is implemented as kernel. Parallel task partition or computing pipeline can be used to speed up complex calculations (as surveyed in (Lopez-Novoa et al., 2015)).
Communicate	Use bitwise operations for checking communication and latency hiding technique for data transfer (Aaby et al., 2010).
Non-communicate	Data partitioning on agent's address space (as surveyed in (Lopez-Novoa et al., 2015)).

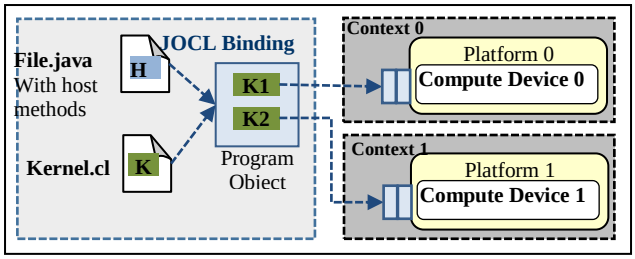


Figure 1. The architecture of the OpenCL execution model with Java code.

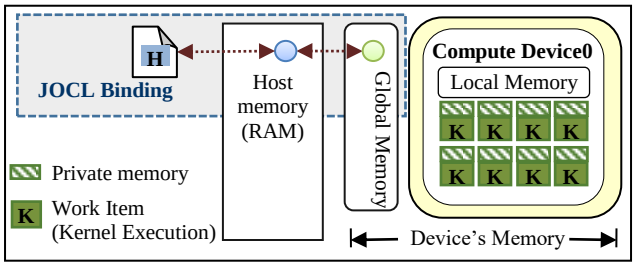


Figure 2. OpenCL memory mapping architecture.

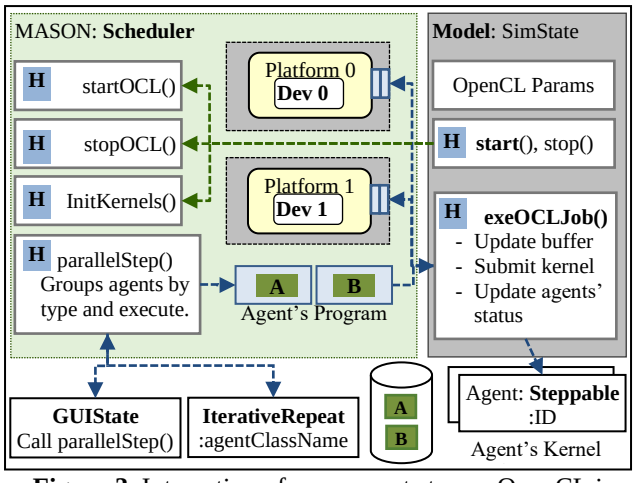


Figure 3. Interaction of components to run OpenCL in MASON's models.

In this paper we propose a technique to classify multiple OpenCL kernels for heterogeneous devices to obtain sustainable overall speedup. Our main contributions are threefold.

(1) We have developed an agent-parallelisation technique on the MASON framework (Luke, 2015) to accelerate Java-based MAS models developed on multiple OpenCL devices. The technique allows modellers to execute the OpenCL computation side by side with an existing multicore computation.

(2) We present a classification technique implemented in the agent scheduler to guide the scheduler as to which OpenCL device is most suitable for a kernel. Central to the classification is performing static and dynamic profiling on the kernel code and using the information obtained to feed the multiclass support vector machine (SVM) classifier.

(3) We show that using the proposed classification technique in the scheduler, the overall speedup of eight kernels used in our experiment outperform the traditional first-come-first-serve (FCFS) scheduler.

The rest of this paper is organised as follows. The next section introduces a new agent-parallelisation technique. Section 3 presents the proposed classification technique. Section 4 analyses the speedup gained from scheduling eight kernels with the proposed technique in comparison to the FCFS and oracle scheduling counterparts. Finally, Section 5 brings this paper to a conclusion.

2 Accelerating Java-Based MAS using OpenCL

Agent-based simulations (ABS) can be grouped into four classes according to the heterogeneity of the agents and how agents interact with each other (Stone and Veloso, 2000). These classes are: (1) homogeneous non-communicating, (2) heterogeneous non-communicating, (3) homogeneous communicating, and (4) heterogeneous communicating. As shown in Table 1, ABS classes can be parallelised on heterogeneous systems using different techniques. A key technique common to both homogeneous and heterogeneous ABS models is agent parallelisation. In agent parallelisation, agents' behaviours are implemented in separate functions that are ready to be offloaded either to GPUs or other accelerators. In OpenCL, these functions are called kernels. A kernel can be scheduled in parallel to be executed on any OpenCL supported device.

Most existing MAS models are implemented on legacy simulation frameworks, *e.g.*, MASON, Repast, FLAME, JADE or NetLogo (as reviewed in (Marurngsith, 2014; Parry and Bithell, 2012; Railsback et al., 2006)). Many of these frameworks have been developed in Java (MASON, Repast, JADE) or Scala

(NetLogo). (Ho et al., 2015) successfully modified the MASON library to support CUDA GPU computing. They achieved a speedup of 187x using the JCUDA binding. However, the target accelerator for CUDA computing is limited to the Nvidia GPUs. The next subsection presents an agent parallelisation alternative technique for OpenCL accelerators.

2.1 The OpenCL Execution Model

The architecture of the OpenCL execution model in Java using the JOCL binding (JOCL, 2011) is shown in Figure 1. An OpenCL project comprises a host application and kernel functions. The execution environment is set by the host application in four steps: (1) getting the number of OpenCL platforms available in a machine, (2) getting the number of available devices for each platform, (3) creating a work space (context) for each platform, and (4) creating a job submission queue (command queue) for each device. The host application also prepares tasks (kernels) ready to be offloaded to available devices. The latter process involves creating a program object, reading kernel files (.cl) and invoking the OpenCL compiler to create binaries for all the kernel functions.

The host application allocates memory buffers in the machine's main memory (called host memory), and also manages data transfers and memory-address mapping between the host memory and the devices' memory (see Figure 2). The memory buffers are used to transfer data to/from a device.

2.2 Implementing OpenCL-Enabled MAS on MASON

The components used to implement the OpenCL-enabled MAS on MASON are shown in Figure 3. Three classes of the MASON simulation engine were modified: Scheduler, GUIState and IterativeRepeat. Four methods
were also added to the Scheduler class to perform the tasks of the host application. The handlers of OpenCL's context, devices, command queues, program, and kernels are all kept in the Scheduler. These handlers can be accessed from the model. This model is a Java class derived from the SimState class which should be manually modified by the modeller. The model accelerated in this work, the Student Schoolyard Cliques model, has been taken from the MASON tutorial. This model consists of student agents and one anonymous agent. Users can enable OpenCL support and select the OpenCL devices via parameters. If the OpenCL flag is set, the start method will invoke a method in the Schedule class to initialise the OpenCL environment. The start method also allocates host memory buffers and maps them to the memory of the OpenCL devices. When a kernel is invoked, data in the host memory buffers is transferred to the memory of the device associated with the kernel.

We use an agent-parallelisation technique to accelerate the execution of student's behaviour at every time step. Our technique consists of rewriting the behaviour of a Student agent as kernel function in OpenCL. Therefore, if the OpenCL flag is set, at every simulation step, the kernel implementing the agents' behaviour is scheduled to be executed on a suitable OpenCL device. If the OpenCL flag is off, the original scheduling method of the MASON framework, called step, is invoked to perform multicore execution. To facilitate the OpenCL scheduling process, we added a method, called parallelStep, to the MASON scheduler. Similar to the original step method, the parallelStep method gathers objects based on their timestamp and puts them into a list. However, the parallelStep method differs from the step method in that the former method splits objects into two lists, an execution list for CPUs and another list for OpenCL devices. In this process, the IterativeRepeat class is used to get objects' class names that are checked against the list of kernels' class names. When a match is found, the object is removed from the CPU list and added to the list of objects for OpenCL devices. Objects in the CPU list are executed using traditional multicore execution. Simultaneously, the objects for OpenCL devices are scheduled on the OpenCL command queue handled by the model. We implemented the method executeOpenCLJob in the model to carry out kernel submission. The executeOpenCLJob method updates memory buffers, submits kernels to the command queue and reads results back from OpenCL devices before updating the status of all agents accordingly. This parallel scheduling process is iterated until the end of the simulation. It is important to note that the scheduler uses the kernel classification to identify which command queue (OpenCL device) is suitable to execute a kernel (see Figure 4).

3 OpenCL kernel classification

The technique for kernel classification has been modified from (Wen et al., 2014) to identify which device is suitable for executing a kernel. In (Wen et al., 2014) the binary SVM classifier gives a more accurate prediction than that of the neuron network technique. The binary SVM classifier was used to classify kernels into low and high speedup groups to identify as a suitable device either a CPU or a GPU. Nevertheless, after collecting the execution time of 21 workloads for training the classifier, we noticed that the execution time obtained from two different GPUs could be significantly different. Thus, in this work, we adopted the multiclass SVM classification technique (Chih-Wei and Chih-Jen, 2002) to support selection from more than two OpenCL devices. Kernels are classified into k groups, where k is the number of available OpenCL devices so that guided results can be used not only with a GPU or CPU, but also with specific devices. The connection of the proposed SVM kernel classifier with the MASON framework is shown in Figure 4.

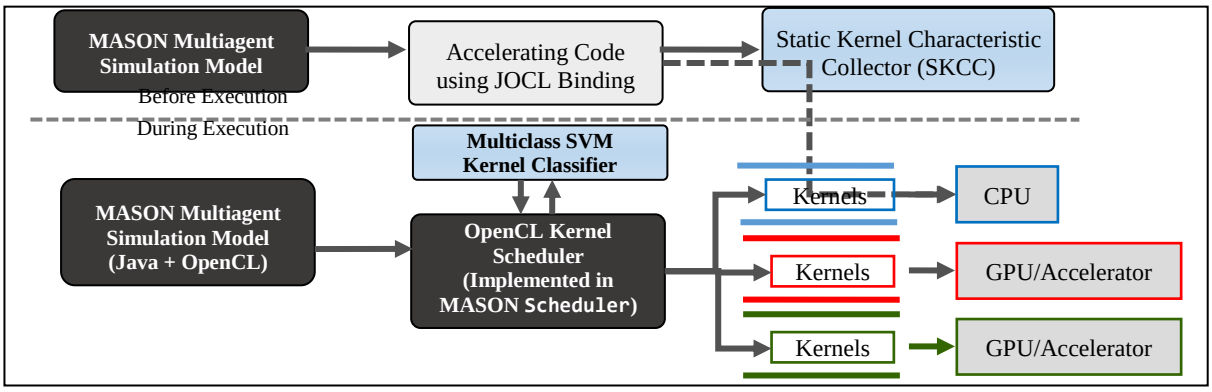


Figure 4. Overall connection of the proposed SVM kernel classifier with the MASON agent-based simulations to offload agents to OpenCL devices.

3.1 Workload Characteristics Profiling

We captured fifteen features of host and kernel code from both a static profiling tool and by inserting profiling functions into the original code. The features collected are shown in Table 2. Twenty-one workloads (available on the Nvidia and Intel websites) were used for training the classifier (see Table 3). The features of kernels from these workloads were collected and passed to the SVM module for training.

3.2 Support Vector Machine Multiclass

The SVM is a well-known supervised binary classifier. The characteristics of the workloads were used to supervised the classifier. Trained workloads were executed on OpenCL devices available in our target machine (see Table 4) to collect the execution times of the workloads. Each workload was labelled to the device with the fastest execution time.

To allow the classifier to work with more than two classes, the well-known One-against-ALL or One Versus the Rest method (Chih-Wei and Chih-Jen, 2002) was used. First, the number of classes was defined as the number of OpenCL devices available in the system *i.e.*, k classes. Second, the binary classifier was constructed by separating one class from the rest. For example, if k=3, the pair wise of binary classifiers are 0 with (1, 2), 1 with (0, 2) and 2 with (0, 1). After that, training data for the classifier were re-labelled and trained for each possible pair of classes. The results were combined to get a multi-class classification according to the maximum output. Note that the SVM classifier from the Intel Data Analytics Acceleration Library was used in this work.

Table 2. Features used in the classification.

At	**Collected Features** (# = number of)
Host	#iteration, workgroup dimension, global size, local size, input buffer size, output buffer size
Kernel	#parameters, #barriers, #math functions, #int and float scalar operations, #int and float vector operations, #atomic, #control

Table 3. List of workloads used in the experiment.

Workload		**Input size**	**#KN**[a]	**Dim**[b]
BlackScholes	Nvidia	1.1M	2	1
ConvolutionSeparable		150M	4	2
DCT8x8		50M	2	2
DotProduct		25M	1	1
FDTD3d		452M	1	2
HiddenMarkovModel		404	2	2, 1
MatrixMul		128K	1	2
MatVecMul		440M	6	1
MersenneTwister		96M	2	1
Reduction		67M	2	1
Scan		54M	4	1
Transpose		33M	5	2
VectorAdd		137M	1	1
BitonicSort	Intel	134M	1	1
GEMM		188M	1	2
GodRays		61M	1	1
MedianFilter		134M	1	2
ProcGraphicsOpt		8M	3	2
SimpleOptimizations		134M	1	1
ToneMapping		61M	1	2
ToneMappingMultiDevice		122M	1	1

Table 4. Experimental Platform Information.

Detail	**OpenCL**	**Platform**
Host Machine	Dev 0	Intel Core i7-4710HQ (2.50 GHz, 6 MB L3 Cache, up to 3.50 GHz), 8GB RAM
Accelerator	Dev 1	NVIDIA GeForce GTX 850M (4GB GDDR3), 640 cores, 902MHz, Memory of 4096MB, OpenCL1.2
	Dev 2	Intel(R) HD Graphics 4600 (No dedicated Memory, using max of 1.7 GB RAM), 100 Effective SPUs Count, 400MHz, OpenCL1.2
OS		Windows 10 Pro

4 Experimental results and discussion

Two experiments, using three OpenCL devices, were carried out on a Windows-based machine to confirm the performance of the accelerated MASON MAS model and the proposed kernel classification technique. The specification of the experimental platform is listed in Table 4.

4.1 Performance of the Accelerated MAS Model

In our first experiment, we analysed the performance of the accelerated Student Schoolyard Cliques by comparing the OpenCL execution time against the original multicore execution time as baseline. The baseline execution of the model using one thousand to a quarter of a million students is shown in Figure 5. The execution time shows a linear growth in agents.

The OpenCL-enable model was executed on the Nvidia GPU (Dev1) and on the host CPU (Dev0). We collected the execution time and calculated the speedup relative with the baseline performance (see Figure 6). The results show a similar speedup obtained from two computing devices. However, when the number of agents (students) is small, the OpenCL execution is slower than the multicore execution (1,000 to less than 5,000 agents). Furthermore, the accelerated model outperforms the original model when there are more than 5,000 agents. The highest speedup obtained is 27.6x faster with a constraint of 250,000 agents.

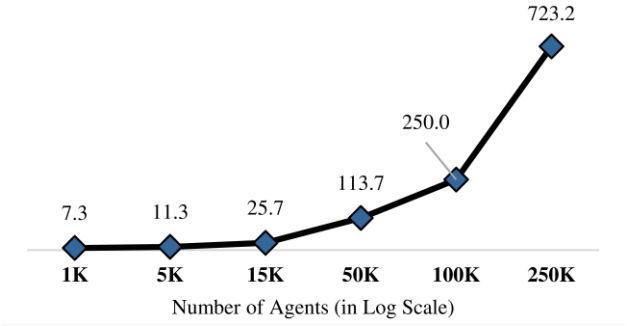


Figure 5. Baseline execution time (in sec.) of the Students model on CPU.

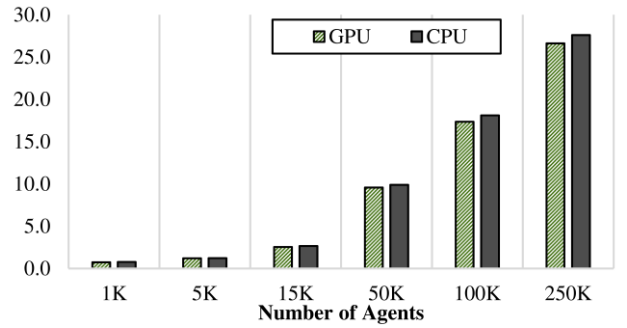


Figure 6. Speedup of the OpenCL execution over the baseline.

Certainly, the available memory space was an issue for the accelerated MAS models on the MASON framework. This is because, in the framework, many key data are defined as double *e.g.*, the coordinate representing agent's position. In our accelerated model, agents resided on a 2D continuous space. Each agent requires 128 bits to store the x, y coordinate. Therefore, the size of the input and output buffers grows with the number of agents. The memory available on our experimental platform reached its limit of devices at 250,000 agents (students). It is also important to note that the double data type is not supported in some OpenCL devices. Consequently, compatibility of the available devices must be verified before executing OpenCL code.

4.2 Performance of the Classifier

The aim of our second experiment was to quantify the effectiveness and accuracy of the classifier. We used the classifier with accelerated MAS kernels. However, MAS models can only generate a limited number of kernels that might not be representative of the key computation load of general kernels. Consequently, only eight kernels and twenty-one scientific workloads were used to test the SVM classifier. The list of kernels used in this experiment and their execution times (in millisecond) measured on the OpenCL devices of the target platform are shown in Table 5. The BitonicSort workload shows the longest execution time, and the most aggressive acceleration on Dev1 (the Nvidia GPU). Note that the execution time obtained from two different GPUs (Dev1 and 2) are very different.

A classification of kernel features is shown in Table 5. The output prediction is used for scheduling the kernels combining the FCFS technique with the results of the classification. The results of all four scheduling techniques, oracle scheduling, the proposed method (FCFS + Classification), all kernels scheduled on the fastest device, and FCFS, are shown in Figure 7 – Figure 10, respectively. The obtained results show that in terms of overall execution time (in milliseconds), the proposed method performs similarly to the fastest device technique, but is 25% slower than oracle scheduling. Scheduling based on the proposed method outperforms the FCFS-only technique by over 45%. In this experiment, suitable devices for seven kernels were correctly identified but one failed. The classifier identified an incorrect device for kernel B (Median Filter) *i.e.*, Dev2 was selected instead of Dev0. Thus, the prediction accuracy of our proposed classifier is 87.5%. However, this accuracy rate cannot be generalised yet as the number of tested kernels was small.

The classifier was also used with the kernels created in the accelerated Student Schoolyard Clique model. In this case, the classifier identified the correct device. However, as the model has only one type of

kernel, a very small speedup is observed when using the FCFS + Classification method. Moreover, the kernel execution time on CPU and GPU is very similar (see Figure 6). Thus, more accelerated MAS models should be used to confirmed the effectiveness of the classifier on the MAS acceleration.

Table 5. Kernel Execution Time (in Milliseconds).

Kernels		Classify	Dev0	Dev1	Dev2
A	VectorAdd	Dev2	0.04	4.33	0.02
B	MedianFilter	Dev2	14.29	34.39	35.91
C	BitonicSort	Dev1	4,641.23	2,709.99	4,834.55
D	SobelGraphic	Dev1	3.42	1.26	1.98
E	MT Naïve	Dev1	483.36	198.31	602.09
F	MT Simple Copy	Dev1	377.40	135.43	215.87
G	MT Shared Copy	Dev1	483.48	135.43	305.85
H	Matrix Transpose	Dev1	575.98	138.60	319.77

5 Conclusions and future work

In this paper, an OpenCL-kernel classification technique to identify a suitable OpenCL device for a kernel, has been proposed. An agent-parallelisation technique for multiple OpenCL devices to accelerate a JAVA-based MAS model on the MASON framework has been discussed. A modified SVM for multiclass classification has been used to guide the scheduler to offload kernels to suitable devices. The proposed classification could achieve 87.5 percent of accuracy on tested workloads.

The accelerated Java MAS achieved a 27x speedup in comparison to the original multicore execution using a maximum of 250K agents. The proposed classifying kernel technique arranged the MAS kernel correctly as is demonstrated by scheduling eight different computational kernels. The results show that our classification technique is slower than oracle scheduling, but outperforms FCFS scheduling. The latter suggests that kernel classification can be an alternative option for sustainable speedup in accelerated Java-based MAS models. However, in order to achieve an effective scheduler in the MAS engine, future work must focus on classifying kernels that has been generated from a wider variety of MAS models.

Acknowledgements

We thank the reviewers for their valuable comments. We thank contributors to the MASON, JOCL and OpenCL community website for discussion and lesson learned. We thank Professor Roland Ibbett and JC Diaz Carballo for improving the readability of this paper.

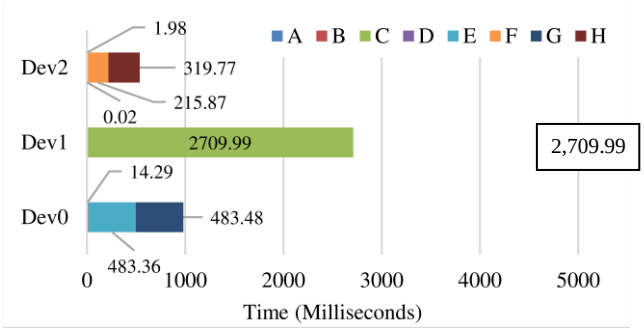


Figure 7. Oracle Scheduling.

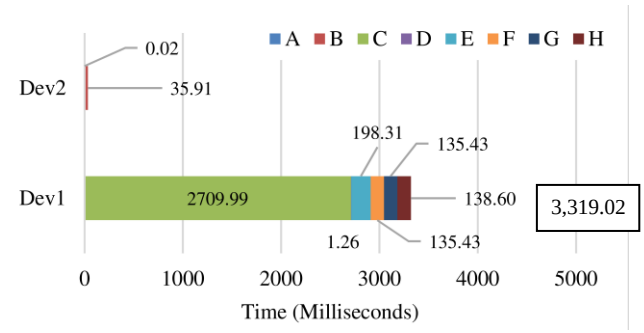


Figure 8. FCFS + Classification Scheduling.

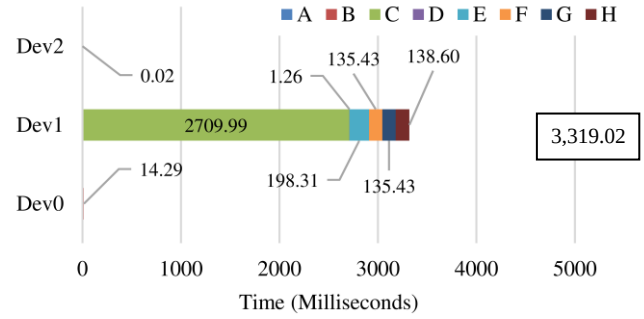


Figure 9. All kernels scheduled on the fastest device.

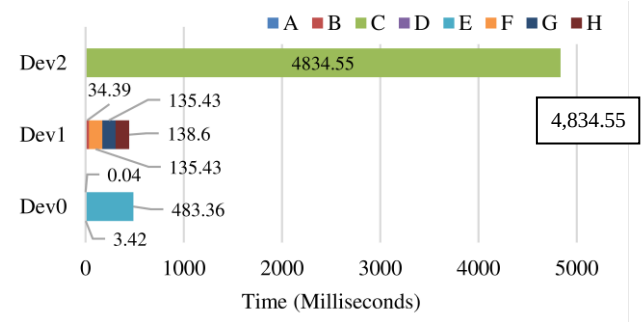


Figure 10. FCFS Scheduling.

References

Brandon G. Agby, Kalyan S. Perumalla, and Sudip K. Seal. Efficient simulation of agent-based models on multi-GPU and multi-core clusters. In *the 3rd International ICST Conference on Simulation Tools and Techniques*, SimuTools, 2010.

AMD Developer Central. *APARAPI: An Opensource API for Expressing Parallel Workloads in Java*. Available

via http://developer.amd.com/tools-and-sdks/opencl-zone/aparapi/, 2011.

R. Dolbeau, F. Bodin and G. C. de Verdiere. One OpenCL to rule them all? In Proceedings of *Multi-/Many-core Computing Systems (MuCoCoS), 2013 IEEE 6th International Workshop on*, 2013.

Khronos Group. *OpenCL - The open standard for parallel programming of heterogeneous systems*. Available via http://www.khronos.org, 2013.

A. Hayashi, M. Grossman, J. Zhao, J. Shirako and V. Sarkar V. Accelerating Habanero-Java programs with OpenCL generation. In Proceedings of *ACM International Conference Proceeding Series*, 2013.

N. M. Ho, N. Thoai and W. F. Wong. Multi-agent simulation on multiple GPUs. *Simulation Modelling Practice and Theory, 57*: 118-132, 2015.

Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks, 13*(2): 415-425, 2002.

JOCL. *jocl.org: Java bindings for OpenCL*. Available via http://www.jocl.org/, 2011.

X. Li, W. Cai and S. J. Turner. Supporting efficient execution of continuous space agent-based simulation on GPU. *Concurrency Computation*, 2016.

U. Lopez-Novoa, A. Mendiburu and J. Miguel-Alonso. Survey of performance modeling and simulation techniques for accelerator-based computing. *IEEE Transactions on Parallel and Distributed Systems, 26*(1): 272-281, 2015.

Sean Luke. *Multiagent Simulation and the MASON Library*. Available via https://cs.gmu.edu/~eclab/projects/mason

Worawan Marurngsith. Computing platforms for large-scale multi-agent simulations: The niche for heterogeneous systems. *Vol. 8669 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 424-432, 2014.

A. Matoga, R. Chaves, P. Tom and N. Roma. A flexible shared library profiler for early estimation of performance gains in heterogeneous systems. In Proceedings of *High Performance Computing and Simulation, HPCS*, 2013.

R. Mokhtari and M. Stumm. BigKernel - High performance CPU-GPU communication pipelining for big data-style applications. In *Proceedings of The International Parallel and Distributed Processing Symposium, IPDPS*, 2014.

Hazel R. Parry and Mike Bithell. Large Scale Agent-Based Modelling: A Review and Guidelines for Model Scaling. In *Agent-Based Models of Geographical Systems*, pages 271-308, 2012.

Steven F. Railsback, Steven L. Lytinen and Stephen K. Jackson. Agent-based Simulation Platforms: Review and Development Recommendations. *SIMULATION, 82*(9): 609-623, 2006.

C. J. Rossbach, Y. Yu, J. Currey, J. P. Martin and D. Fetterly. Dandelion: A compiler and runtime for heterogeneous systems. In *Proceedings of the 24th ACM Symposium on Operating Systems Principles, SOSP*, 2013.

Rafael Sachetto Oliveira, Bernardo Martins Rocha, Ronan Mendonça Amorim, Fernando Otaviano Campos, Wagner Meira, Jr., Elson Magalhães Toledo and Rodrigo Weber Santos. Comparing CUDA, OpenCL and OpenGL Implementations of the Cardiac Monodomain Equations. In *Parallel Processing and Applied Mathematics*, Vol. 7204, pages 111-120, 2012.

K. Sato, K. Komatsu, H. Takizawa and H. Kobayashi. A History-Based Performance Prediction Model with Profile Data Classification for Automatic Task Allocation in Heterogeneous Computing Systems. In *Proceedings of Parallel and Distributed Processing with Applications, ISPA*, 2011.

Peter Stone and Manuela Veloso. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots, 8*(3): 345-383, 2000.

Y. Wen, Z. Wang and M. F. P. O'Boyle. Smart multi-task scheduling for Open CL programs on CPU/GPU heterogeneous platforms. In Proceedings of *2014 21st International Conference on High Performance Computing, HiPC*, 2014.

Y. Yan, M. Grossman and V. Sarkar. JCUDA: A programmer-friendly interface for accelerating java programs with CUDA. *Vol. 5704 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 887-899, 2009.

Experiences and Trends in Control Education: A HiOA/USN Perspective

Tiina M. Komulainen[1] Alex Alcocer[1] Finn Aakre Haugen[2]

[1]Department of Electronics Engineering, Oslo and Akershus University College of Applied Sciences (HiOA), Oslo, Norway, tiina.komulainen@hioa.no, alex.alcocer@hioa.no

[2]Institute of Electrical Engineering, Information Technology and Cybernetics, University College of Southeast Norway (USN), Porsgrunn, Norway, finn.haugen@usn.no

Abstract

Global trends in higher education including e-learning, massive open online courses, and new teaching methods have positively affected control education. Control course content has evolved due to changes in industrial practices and the increasing availability of affordable computer hardware and software. Continuous developments in virtual remote and real laboratories have made hands-on tasks more accessible and affordable. In this article, we share our experiences of undergraduate and graduate control education at the University College of Southeast Norway (USN), and Oslo and Akershus University College of Applied Sciences (HiOA). First, we present an overview of the course content at our institutions, and then, we give examples of the development of real and virtual laboratories, online course materials, new learning platforms, and teaching methods.

Keywords: control education, control laboratories, virtual laboratories, simulation, learning management systems, active learning methods

1 Educational trends

1.1 Trends in higher education

Massive open online courses (MOOCs), e-learning, electronic learning management systems, and student active learning methods have become major trends in higher education in Science, Technology, Engineering, and Mathematics (STEM).

During the past decade, the variety of massive open online courses (MOOC) has expanded and many top universities are offering a wide spectrum of courses (Hansen and Reich, 2015). MOOCs combine teaching from the best academics, modern pedagogy, interactive content, virtual laboratories, and online group discussions delivered through non-profit platforms such as edX, Coursera, and Udacity (Waldrop, 2013). However, the academic content should be supplemented with hands-on experiments supervised by experienced teachers in order to build practical skills (Bartholet, 2013).

For on-campus STEM education, student active learning methods have been proven to increase students' learning outcomes and to decrease drop-out rates (Fraser et al., 2014; Freeman et al., 2014; Hake, 1998). Examples of the successful implementation of student active learning methods in groups in technology-rich rooms are SCALE-UP (Student-Centered Active Learning Environment for Undergraduate Programs) at North Carolina State University (Beichner et al., 2007) and TEAL (Technology-Enabled Active Learning) at the Massachusetts Institute of Technology (Dori and Belcher, 2005). The pedagogy is typically based on Flipped Classroom (FC) methodology, where students are required to have their first exposure to the subject material at home *prior to* class, and where class time is spent *working with* the material (Bergmann and Sams, 2012).

1.2 Trends in teaching aids for control education

Based on the 62 papers presented at the 10th IFAC Symposium on Advances in Control Education (Rossiter, 2013), course development is most active in the following topics: remote laboratories (21%), real laboratories (19%), teaching aids (19%), virtual laboratories (11%), e-learning (11%), robotics (10%), and course content (8%). Many educators aim to make part of the resource and time demanding real laboratories more easily accessible through the internet. However, real laboratories are needed in order to ensure practical hands-on skills for the students.

1.3 Trends in the content of control education

Taking well-known text-books, e.g. (Dorf and Bishop, 2016; Franklin et al., 2014; Nise, 2015; Seborg et al., 2011), as indicators of the course content, it seems that the theoretical content of control courses has not changed much over the last decades. Differential equations, transfer functions, state-space models, and frequency response – in the continuous-time and in the

discrete-time domain, comprise the basis, as they did decades ago. Mathworks MATLAB seems to be the default computing tool upon which exercises in textbooks are based, but National Instruments MathScript and LabVIEW are also used as tools.

We find it somewhat surprising to observe that most textbooks apparently aiming to present a good basis for control theory, do not include model-based predictive control (MPC), with (Seborg et al., 2011) as the exception, despite the fact that MPC theory and applications are frequent topics in journals and conferences, as well as there being many commercial software packages for MPC. One reason for the lack of focus on MPC may be that its theoretical basis is optimization theory – a topic not usually taught at undergraduate level.

2 Control education at HiOA and USN

In this article, we share our experiences of undergraduate and graduate control education at the University College of Southeast Norway (USN), and Oslo and Akershus University College of Applied Sciences (HiOA). First, we present an overview of the course content at our institutions, and then, we give examples of the development of real and virtual laboratories, online course materials, new learning platforms, and teaching methods.

2.1 Control education at USN/Porsgrunn

Subsections are numbered and style *"Heading 2"* should be used.

The University of Southeast Norway (USN) has approximately 16,000 students. Control is taught in various courses at three different campuses. The courses covered here are introductory courses in the bachelor and master programs at the Porsgrunn campus.

The control courses have developed over the years. The main driving forces behind the developments are:

- A desire to increase the students' ability to handle practical control challenges. This requires developing both the pedagogics and the content of the courses.
- Feedback from students, in particular from those who have industrial experience in automation and control.
- Teachers' experience in research and development, in particular the relationship between theory and practice.
- Technological changes entailing increasing availability of affordable computer hardware and software.

In the following, firstly the development of course content is described, and secondly, pedagogical development is described.

Content development

Highlights of the content development are:

- Only experimental PID controller tuning methods are presented, both open loop tuning and closed loop tuning, are taught. Open loop tuning focuses on a process of step-response interpretation of the Skogestad PI tuning rules assuming integrator + transport delay process dynamics (Skogestad, 2003), but also tuning double integrator process dynamics is covered (the double integrator can represent bodies to be position controlled, e.g. ships). Closed loop tuning focuses on the Ziegler-Nichols Ultimate Gain method, both the original tuning rules (Ziegler and Nichols, 1942) and modified tuning rules. Frequency response based tuning methods are not covered.
- Feedforward control with possibly nonlinear differential equation models where the feedforward controller is obtained by substituting the process output variable by its set point and then solving the model for the control variable.
- The Laplace transform, transfer functions, and frequency response analysis are very briefly covered. Down-toning frequency response is in agreement with the low priority given to this topic as indicated by the industrial perspective in the reports (Edgar et al., 2006) and (Haugen, 2009).
- Leaving out theoretical stability analysis in the frequency domain. However, the gain margin and phase margin of control loops are introduced using an experimental loop stability analysis approach (Haugen, 2012).
- Discrete-time algorithms of the PID controller, a time-constant measurement filter, and process simulators.
- Principles and applications of model-based predictive control (MPC) are introduced as the most important model-based controller.
- In one of the introductory courses, an industrial process and control system simulator is introduced (the Kongsberg Oil & Gas Technologies K-Spice simulator).
- Programming skills, making the students able to actually implement control, filter, and simulation algorithms. To this end, National Instruments LabVIEW is introduced as the programming tool.

Pedagogical development

Highlights of the pedagogical developments are:

- Interactive real-time simulators from the SimView library (Haugen, 2012) are used extensively in the theoretical exercises.
- Instructional videos supplementing the lectures (Haugen, 2011).

- During 2016 and 2017, two introductory control courses will be offered both as online courses and traditional campus-based courses. Instructional videos will substitute traditional lectures in the online courses. However, laboratory exercises will, to the extent practical, still be a part of the course, requiring the online students to come to the campus to carry out the experimental work over two or three days.
- A relatively large number of laboratory exercises based on the air heater (Figure 1) are closely integrated with the lectures.

2.2 Control education at HiOA/Oslo

HiOA has approximately 18,000 students, 1,900 study engineering and 310 are undergraduate students in electronics engineering. At the undergraduate level HiOA offers courses in Dynamic Systems, Control Systems I, Control Systems II and Instrumentation. The courses cover the following topics:

Dynamic Systems: Basic introductory course on mathematical modeling and dynamic systems analysis. Differential equations, transfer functions, block diagrams, state-space models, frequency analysis, and time response.

Control Systems I: Basic introductory course on control. PID regulator, process simulation, frequency domain control design, Introduction to multivariable control.

Control Systems II: More advanced topics in control. Noise filtering, System identification, Kalman filtering, LQR/LQG control, MPC control. Introduction to nonlinear control.

Instrumentation: Instrumentation for control system engineers, sensor and actuator specifications, instrumentation diagrams, regulations and safety, PLC architecture and PLC programming.

Industrial hardware and software such as ABB's 800xA control system and Kongsberg's K-Spice simulator, are used in the laboratories for all our control courses.

2.3 Accessible Laboratory Exercises

At USN, a number of laboratory exercises are based on the air heater (Haugen, 2010) shown in Figure 1. Together with LabVIEW on students' laptops and the NI USB-6008 IO device, laboratory exercises are run throughout the course, with students working in groups of two or three, see Figure 2. Twenty-six identical rigs have been constructed in-house.

The laboratory assignments cover:

1. Manual temperature control, monitoring, and data logging to file.
2. Implementation of a dynamic process simulator from a time-constant and time-delay model with default model parameter values.
3. Adaptation of the mathematical model, i.e. parameter estimation, using a straightforward, "brute force" least squares method implemented in nested for-loops.
4. Implementation of a discrete-time PI controller and an on/off controller.
5. Implementation of a discrete-time time-constant lowpass filter.
6. Controller tuning using Skogestad's tuning rules and the Ziegler-Nichols Ultimate Gain method, see above.
7. The stability of the control loop. Hitherto, a qualitative analysis is included, including the stability impact of controller gain (both absolute value and sign), integral time, and filter time-constant. In the future, an experimental estimation of gain margin and phase margin [20] will be included.
8. Experimental, table-lookup feedforward control with air flow (disturbance) measurement as input signal and heater control signal as output signal.
9. Temperature control with an industrial PID controller (Fuji PGX5), instead of the LabVIEW-based control system.

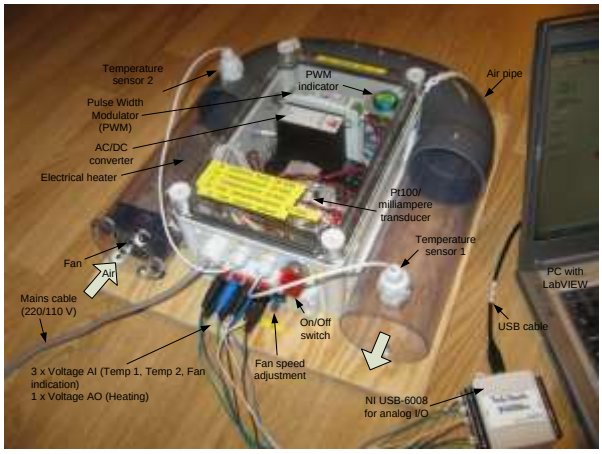


Figure 1: Air heater laboratory rig for temperature control. The voltage control signal manipulates the power delivered by the electrical heater. The outlet temperature is measured by a Pt100 element. The air flow through the pipe can be manually adjusted, representing a (measured) process disturbance.

Figure 2: Students working on laboratory assignments in groups.

2.4 Virtual laboratories / Commercial Large-Scale Simulators

In order to familiarize our students with industrial tools, and to give them insight into chemical processes, commercial large-scale dynamic process simulators have been utilized at HiOA (Komulainen and Løvmo, 2014; Komulainen, 2013; Komulainen et al., 2012). The simulation modules have been developed using the didactic model and the simulator training structure: briefing (lecture) – simulation (guided virtual laboratory) – debriefing (workshop). The simulation software K-Spice is provided by Kongsberg Oil and Gas Technologies Figure 3.

In the following, an example is given of the Dynamic Systems course which is taught to about 60 second year undergraduate electronics engineering students. Two of the learning outcomes of the course are "Student can characterize responses of first and second order systems in time and frequency domain" and "Student can carry out simulation of dynamic systems and interpret the results". The goal of the simulation module is to give the students hands-on skills to use an industrial simulator, to make a step change and identify the process response. The parameters of the process response will be used further for control tuning purposes.

The experiences from the simulator module are positive, the students and the teacher were very positive in their evaluation, 97% of students agreed that simulation exercises increase their understanding of process dynamics. However, the final exam results for the identification tasks were lower than the average final exam mark for both 2013 and and 2014 (Komulainen and Løvmo, 2014). In order to enhance learning through simulation training, we are currently working on developing an automatic assessment system (Marcano and Komulainen, 2016).

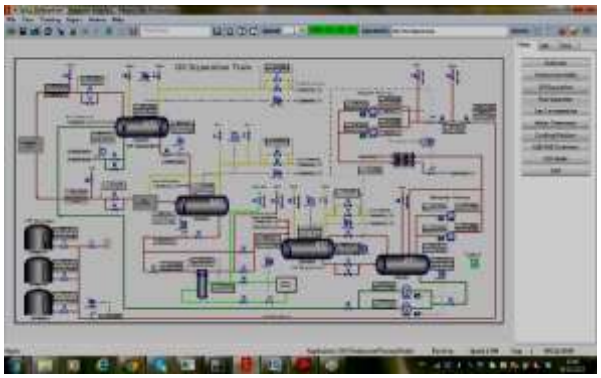


Figure 3: K-Spice® generic oil and gas production simulator.

2.5 Jupyter notebooks and interactive code

Numerical simulation tools have a crucial role in increasing the understanding of control theoretical concepts as well as providing insight and promoting the curiosity and engagement of students (Dormido et al., 2005; Grega, 1999). Typically, MATLAB/Simulink is the numerical simulation software tool of choice in most current control systems courses. Alternatives exist that are gradually providing similar functionalities, which are also open source and free. These include GNU Octave (Eaton, 2016) and Python.

Automatic control is a highly multidisciplinary subject, which has been referred to as the "hidden technology"(Åström, 1999). It involves, among others, the fields of mathematics, physics, electrical and mechanical engineering. In practice, all modern control systems are eventually implemented using some sort of software and programming language. Software development is therefore becoming an increasingly important and required skill, and its importance has naturally gradually increased in control engineering course curriculums (Bencomo, 2004; Åström and Kumar, 2014).

A relatively recent technology enables interactive code to be integrated with rich text in so-called notebooks (Shen, 2014). Notebooks can be viewed and executed using a simple internet browser. This provides an excellent way of distributing educational content and providing students with an initial executable code with which to experiment and develop new ideas. Jupyter is at the forefront of this technology and provides support for a great number of programming languages including Julia, Python, and R (Project Jupyter, 2016). Notebooks can be viewed in an internet browser using a notebook viewer (nbviewer) which does not require any special software. Additionally, the students can chose to download the notebooks to their computers where they have the possibility to interact and modify the initial code.

Python is a popular object-oriented scientific programming language that is becoming increasingly used in research and industry. Several Python libraries exist that are of interest to control engineering students.

For instance numpy and matplotlib provide numerical and data visualization tools that are quite similar to MATLAB. The python control systems library (Murray and Livingston, 2009) is particularly interesting. It implements basic operations for analysis and design of feedback control systems including block diagram algebra, Bode and Nyquist plots, time response, etc. By installing a Python scientific distribution, such as continuum analytics anaconda, the student can easily experiment with these open source tools at no cost. An example of notebook using python, numpy, and matplotlib to easily visualize simulation results (Alcocer, 2016).

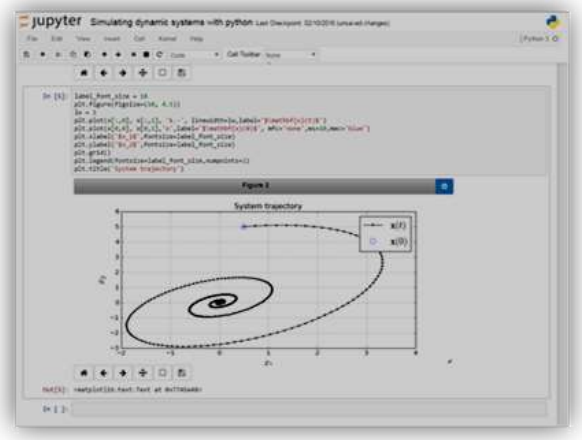


Figure 4: Example of interactive code using a browser, Jupyter notebook, and Python-Control toolbox.

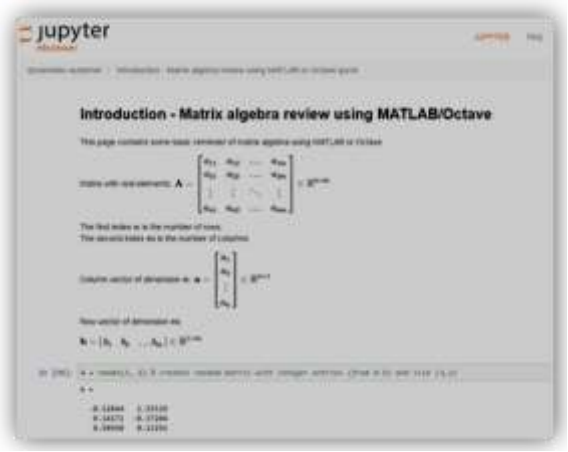


Figure 5: Example of Jupyter notebook using an Octave kernel.

Another interesting possibility is the use of Octave kernel together with Jupyter notebooks. GNU Octave is an open source scientific programming language with a syntax very similar to MATLAB. This provides the possibility of distributing educational notebooks with text, mathematical equations, and code. See Figure 5 for an example of a Jupyter notebook using Octave.

2.6 Learning Management Systems and OpenEdx

OpenEdx is currently one of the most popular open source MOOC platforms. The introductory undergraduate dynamic systems course at HiOA is going to experiment with the use of OpenEdx, see Figure 6. One of the most appealing functionalities is its ability to provide quizzes for the students for each of the units, which provides feedback on and interactivity with the learning experience. With OpenEdx, it is simple to include LaTeX style mathematical expressions integrated in quizzes, which provides a great level of flexibility.

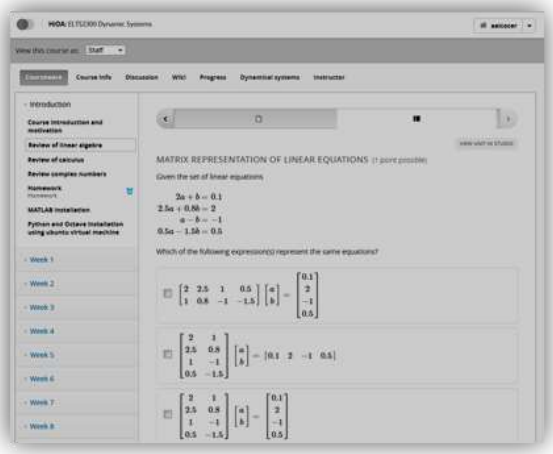


Figure 6: Example of OpenEdx course with quizzes containing mathematical expressions.

2.7 Student Active Learning Methods

At HiOA we have tested Flipped Classroom inspired teaching methods in a technology-rich group room (Komulainen et al., 2015). The experiment was conducted in a dynamic systems course with about 60 students during fall semester 2014. The main goal of the research was to find out if students' learning outcome would increase as a result of the use of student active learning methods. The data collection included students' course evaluation, students' attendance, students' pre and post scores from the Control Systems Concept Inventory (Bristow et al., 2012), teachers' classroom-activity log, five in-class mini-tests, and final exam grades.

The students were given reading assignments with theory quizzes prior to the classroom sessions. During the classes, the students worked in small groups of three to four students and used a small screen at the end of each table to present the work of their group. Short tasks (5-20 min) were given on concepts, theory and basic calculations, long exercises (20-45 min) on modeling of dynamic systems and simulation of these models with Matlab/Simulink. After each task, the teacher chose one of the groups to present its results to the whole class.

These plenary presentations were facilitated with large screens using AirMedia software. Every other week, the students took a 20-minute mini-test on theory and modeling. The mini-test was graded by the peer students immediately afterwards based on the solution presented by the teacher on the SmartBoard.

The students' course evaluation indicated that 70% of the students preferred the active learning classroom to traditional lecturing. Students valued the mini-tests as a tool to monitor their own progress in the course and they emphasized the good learning outcome of the group work. The students gave the course a final average mark of B.

Student attendance of 72% was considered good and above average for this student cohort. However, only 42% of the students answered the quizzes prior to the classes. Students' conceptual understanding increased during the course, the normalized gain was 20% measured by the Control Systems Concept Inventory. The average final grade for the course in dynamic systems was compared to the average final grade for the course in electrical circuits between cohort 2013 (traditional lecturing) and cohort 2014 (active learning methods). The average grade in electric circuits was 3.94 for cohort 2013 and 3.35 for cohort 2014, indicating that cohort 2013 was academically stronger than cohort 2014. However, the difference between the cohorts had become non-significant after the dynamic systems course; the final grade was 2.64 for cohort 2013 and 2.63 for cohort 2014. Although the results were not conclusive, the results indicate that active learning methods applied in 2014 were more valuable to student learning than traditional lecturing.

3 Discussion and conclusions

Global trends in higher education, online course materials and affordable hardware and software have provided great possibilities for making control education more accessible, efficient, and interesting to students, teachers and universities. In this article we have provided examples of experiences at USN and HiOA, and have shown how some of these teaching tools have been applied in control systems courses. Special attention is given to an experiment involving Flipped Classrom methodology together with a technology-rich group room. This teaching methodology was tested with positive results during an undergraduate dynamic systems course. The paper also discusses, among other things, the use of accessible laboratories, industrially relevant virtual laboratories, open source simulation tools, open learning management systems, and new teaching methods that are promising or have been successfully implemented in control systems courses at USN and HiOA.

References

A. Alcocer. *Dynamiske-systemer*. Available via https://github.com/aalcocer/dynamiske-systemer [accessed 18.05.2016, 2016].

J. Bartholet. MOOCs- Hype and Hope. *Scientific American* 309:53-61, 2013. doi: 10.1038/scientificamerican0813-53

R. J. Beichner, J. M. Saul, D. S. Abbott, J. J. Morse, D. L. Deardorff *et al.*, The Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) Project, vol. 1, Research-Based Reform of University Physics, E. F. Redish and P. J. Cooney, Eds., College Park: American Association of Physics Teachers, 2007. [Online]. Available: http://www.compadre.org/per/per_reviews/volume1.cfm. Accessed on 01.02.2015.

S. D. Bencomo. Control learning: present and future. *Annual Reviews in control*, 28(1):115-136, 2004.

J. Bergmann and A. Sams. In *Flip Your Classroom: Reach Every Student in Every Class Every Day*. International Society for Technology in Education, 2012.

M. Bristow, K. Erkorkmaz, J. P. Huissoon, S. Jeon, W. S. Owen, S. L. Waslander, and G. D. Stubley. A Control Systems Concept Inventory Test Design and Assessment. *IEEE Transactions on Education*, 55(2):10, 2012. doi: 10.1109/TE.2011.2160946

R. Dorf and R. H. Bishop. In *Modern Control Systems*. 12 ed. Pearson, 2016.

Y. J. Dori and J. Belcher. How Does Technology-Enabled Active Learning Affect Undergraduate Students' Understanding of Electromagnetism Concepts? *The journal of the learning sciences*, 14(2):243-279, 2005.

S. Dormido, S. Dormido-Canto, R. Dormido, J. Sánchez, and N. Duro. The role of interactivity in control learning. *International Journal of Engineering Education*, 21(6):11-22, 2005.

J. W. Eaton. *GNU Octave*. Available via https://www.gnu.org/software/octave/ [accessed 18.05.2016, 2016].

T. F. Edgar, B. A. Ogunnaike, J. J. Downs, K. R. Muske, and B. W. Bequettee. Renovating the undergraduate process control course. *Computers & Chemical Engineering*, 30:1749-1762, 2006.

G. Franklin, J. Powell, and A. Emami-Naeini. In *A. Feedback Control of Dynamic Systems*. 7 ed. Pearson, 2014.

J. M. Fraser, A. L. Timan, K. Miller, J. E. Dowd, L. Tucker, and E. Mazur. Teaching and physics education research: bridging the gap. *Reports on Progress in Physics*, 77(3):17, 2014. doi: 10.1088/0034-4885/77/3/032401

S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M.P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, *Proceedings of the National Academy of*

Sciences of the United States of America (PNAS), p. 6, 15.04.2014. doi: 10.1073/pnas.1319030111

W. Grega. In Hardware-in-the-loop simulation and its application in control education. *Frontiers in Education*, San Juan, Puerto Rico 1999, volume 2, pages 12B6-7.

R. R. Hake. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66:64-74, 1998. doi: http://dx.doi.org/10.1119/1.18809

J. D. Hansen and J. Reich. Democratizing education? Examining access and usage patterns in massive open online courses. *Science*, 350(6265):1245-1248, 2015. doi: 10.1126/science.aab3782

F. A. Haugen. (2009) Industrifolks syn på automatiseringsutdanningen (in English: Industrial perspective on control education). *AMNytt*. Available: http://techteach.no/publications/amnytt/web

F. A. Haugen. *Lab Station: Air Heater*. Available via http://home.hit.no/¨finnh/air_heater [accessed 18.05.2016, 2010].

F. A. Haugen. *TechVids*. Available via http://techteach.no/techvids [accessed 18.05.2016, 2011].

F. A. Haugen. The Good Gain method for simple experimental tuning of PI controllers. *Modeling, Identification, and Control*, 33(4):141-152, 2012. doi: 10.4173/mic.2012.4.3

F. A. Haugen. *SimView*. Available via http://techteach.no/simview [accessed 18.05.2016, 2012].

T. Komulainen and T. Løvmo. In Large-Scale Training Simulators for Industry and Academia. *55th Conference on Simulation and Modelling*, Aalborg, Denmark, 2014, volume 128-137.

T. M. Komulainen. In Integrating commercial process simulators into engineering courses. In A. Rossiter, editor, *10th IFAC Symposium Advances in Control Education*, University of Sheffield, 2013, volume 10, pages 274-279 Sheffield, 2013. doi: 10.3182/20130828-3-UK-2039.00007

T. M. Komulainen, C. Lindstrøm, and T. A. Sandtrø. Erfaringer med studentaktive læringsformer i teknologirikt undervisningsrom. (in Norwegian), *Uniped*, 8(04):364-372, 2015.

T. M. Komulainen, R. Enemark-Rasmussen, G. Sin, J. P. Fletcher, and D. Cameron. Experiences on dynamic simulation software in chemical engineering education. *Education for Chemical Engineers*, 7(4):e153-e162, 2012. doi: 10.1016/j.ece.2012.07.003

L. A. Marcano and T. M. Komulainen. Constructive Assessment Method for Simulator Training. *In Proceedings of the 9th Eurosim Congress on Modelling and Simulation*, Oulu, Linköping University Press, 2016.

R. Murray and S. C. Livingston. *Control Systems Library for Python*. Available via https://github.com/python-control/ [accessed 18.05.2016, 2009].

N. S. Nise. In *Control Systems Engineering*. 7 ed. Wiley, 2015.

Project Jupyter. *Jupyter*. Available via http://jupyter.org/ [accessed 18.05.2016, 2016].

A. Rossiter. In Proceedings of the 10th IFAC Symposium Advances in Control Education. *In Proceedings of 10th IFAC Symposium Advances in Control Education*, Sheffield, Great Britain, 2013, volume 10. doi: 10.3182/20130828-3-UK-2039.00007

D. E. Seborg, D. A. Mellichamp, T. F. Edgar, and F. J. D. III. In *Process dynamics and control*. international student version 3 ed. Wiley, p. 528, 2011.

H. Shen. Nature toolbox: Interactive notebooks: Sharing the code. *Nature*, 515:151-152, 2014. doi: 10.1038/515151a

S. Skogestad. Simple Analytical Rules for Model Reduction and PID Controller Tuning. *Journal of Process Control*, 13:291-309, 2003.

M. M. Waldrop. Education online: The virtual lab. *Nature*, 499:268-270, 2013. doi: 10.1038/499268a

J. G. Ziegler and N. B. Nichols. Optimum Settings for Automatic Controllers. *ASME*, 64:759-768, 1942.

K. J. Åström. Automatic control: The hidden technology. In P. M. Frank, editor, *Advances in control: Highlights of ECC'99* pages 1-28: Springer Verlag, 1999

K. J. Åström and P. R. Kumar. Control: A perspective. *Automatica*, 50(1):3-43, 2014.

Challenges and New Directions in Control Engineering Education

Kai Zenger

Department of Electrical Engineering and Automation, Aalto University, Finland, Kai.Zenger@aalto.fi

Abstract

The paper discusses the changes and challenges in the current teaching of Automatic Control systems. Modern society has developed into a phase where the traditional process industry is not at all the only area where dynamic modelling, understanding the feedback, control engineering, autonomous systems and generally the discipline of Automatic control have to be mastered. That gives a huge challenge to the teaching of automatic control in general, especially when fewer and fewer students are entering engineering schools and as the basic skills in mathematics and physics seem to be decreasing everywhere. On the other hand, automation (to be understood broadly including automatic control and control engineering, autonomous systems etc.) as a discipline is in a state of change: it seems to be hidden in other engineering fields, and there seems to be opinions that it should actually be taught within specific application areas, e.g. in electrical engineering, machine design, chemical process engineering etc. In the old school of control engineering the idea is actually vice versa: automatic control is seen as a general, mathematically and physically well-defined discipline, which can the be applied in various application areas and engineering fields. The societal and industrial viewpoints must both be considered, when looking at the future of control education. These aspects are discussed in the paper.

Keywords: education, automatic control, autonomous systems, control engineering, curriculum

1 Introduction

Control engineering, control theory or *system theory* are the cornerstones of automatic control or autonomous systems in general. The interdisciplinary nature of control and applications is shown as an example case in Figure 1, where the classical idea of control theory serving a multitude of application areas is demonstrated, (Zenger, 2007).

The above age old idea has been good and well-serving for a long period of time, but today in the modern society there are aspects that suggest a change. Firstly, control has always been considered a difficult topic for the students to learn, and this attitude is getting stronger as the mathematical skills of students are generally considered weaker than before (Rasila et al., 2007). Secondly, the status of automation is perceived weaker, as there is a trend to consider it a part of other well-established engineering fields only, and not a research field of its own right. In addition, the application of dynamic modelling, optimization and feedback control for example, has been extended to a much broader field than before. Cyberphysical systems, Information theoretical systems including communication and the Internet, Big Data, Cloud Computing, Internet of Things produce totally new applications, which need control theory and signal processing. In this respect, the ideas and "new" curriculum for control education (see e.g. (Kheir et al., 1996; Murray, 2003)) are already somewhat out of date.

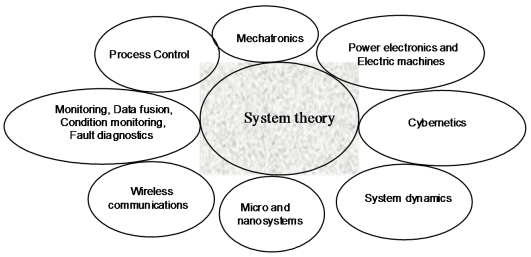


Figure 1. Application fields of control engineering.

The relevant contemporary questions from the education viewpoint are:

- Is automation losing its position as an independent research field and is it going to be combined with other research disciplines?
- The elegant combination of signal processing, control, mechanics etc. is actually hidden in products. Is the value of control not seen anymore?
- The level of teaching automatic control has in general been of high quality. In the industry the graduated students have obtained good positions. How is the situation now and in the future? How can these kinds of issues be "controlled"?
- Should the universities and universities of applied sciences specialize in teaching only some application areas of automation or keep the wide view?
- How is the financing of teaching automation going to develop in the near future?
- What must a graduating student from the automation field know and master now and in the future?

- What about the connections and networking by teachers, students and industrial representatives on a regular basis.

In the paper it is not possible to discuss and solve all the above questions. However, it is important to pose these problems for general awareness and for discussion. That is one target of the paper.

The contents of the paper can briefly be stated as follows. The pedagogical issues, teachers' challenges and comparison to teaching of mathematics are discussed in Section 2. What practical changes and new ideas in control engineering education would be useful to introduce in the modern curriculum in universities and universities of applied sciences is discussed in Section 3. A look into the future is shown in Section 4 and Conclusions are given in Section 5.

2 Control Engineering: The Teacher's Challenge

Control engineering is considered to be a difficult topic to teach. The formalism (given originally by system theory) is based on a firm understanding of basic sciences like mathematics and physics. The background and basic skills in natural sciences of the students vary a lot. It seems that there is a tendency that basic mathematics skills of the students has decreased (Rasila et al., 2007). There are a lot of reports indicating this, but maybe equally valuable is the "touch" of university teachers. According to their opinion the general level of the students in mathematics has decreased considerably in over the past 10-20 years. In engineering schools the attitude of the students in respect to this is usually something like "I have come here to study engineering, not mathematics". Indeed, that is somewhat of a strong argument. Today it seems difficult to persuade the students to accept mathematical rigor by sayting that "the laws of nature have been written in the language of mathematics" or something alike. The youth of today have so much other interests and things to do that they do not want to spend their time in learning mathematics.

A natural extension to the above is to state that teaching control engineering generally as a self-contained theory to serve different applications areas later may not be well understood by the students. Teaching control theory this way is seen as some kind of a new course of mathematics again. Questions and comments like "Why teach everything to everybody?", "Why teach theoretical issues that an engineer does not need in his/her daily work? and "You see - some people do not need these things - ever!" have been asked by the students in numerous occasions where the courses have been evaluated (Zenger, 2007). These questions and the related pedagogical challenge can be analysed using e.g. Ausubel's taxonomy (Ausubel, 1968), which makes a distinction between mechanical, meaningful, assimilative and inventive learning.

It seems logical to assume that this taxonomy is still valid in analyzing the learning process in general. The modern society demands meaningful and inventive learning methods to be used, in order to speed up the studies and get the students to graduate as fast as possible. The old-fashioned teaching methods might be assimilated to more mechanical learning style, which, under the new hypothesis might not be so effective. However, the different students with totally different learning styles are not really consided by this kind of thinking. It is the experience of the author, based on several decades of experience in teaching control systems, that good students learn in spite of teaching, bad students do not learn because they are not motivated, the other students can really be helped by innovative and good teaching practises.

There has been some discussion and doubts about the effectiveness of traditional lectures. The general trend in university level pedagogy researchers seems to be that lectures are not a good teaching method, because they only call for passive learning. It must be admitted that student activation really seems to be (also in Ausubel's taxonomy) the critical point here. Learning by doing seems an effective learning method, if carried out in a good way. But it is not a trivial thing to see, how teaching in this way should really be carried out. If it is done in every course it gives a high work load to the students, who very fast turn critical towards it. Also, there are students who simply like the lectures / exercise hours mechanism. The concept of *Problem-based learning* (Boud and Feletti, 1977) has not been widely used in engineering schools. Surely, applications using the ideas of PBL have been developed and used on laboratory exercises and web courses, see e.g. (Riihimäki et al., 2003; Pohjola et al., 2005).

However, it must always be borne in mind that lectures for a large audience (say, 100–300 students) are really a cheap way to teach. The financiation of teaching is nowadays a very difficult issue.

Comparison of the teaching of control and mathematics is indeed fruitful, because they seem to somehow share similar problems. In basic university level mathematics courses the web-based interactive study material has been tried with success. For example the STACK system has been used (Rasila et al., 2007) to make varying drill problems to the students, who can solve them when and whereever they are, and the results evaluation is really used partly in grading the course. A similar course material was also used in Control education in Helsinki University of Technology (now Aalto University), although in a much smaller scale. It can be concluded that for drill problems the system is effective, but not for real design problems of engineering. On the other hand, computer-based examinations have been discussed a lot lately. The conclusion has so far been that for examinations where only essay kind of answers or very simple drill problems are required, the existing solutions are usable. However, for examinations of more advanced courses with problem solving the existing software is not good enough. For the time being there is a good reason to seek alternative solutions in order to develop teaching.

3 New Curriculum For Teaching Automatic Control

Based on the above there is a good reason to develop new ideas and methods to be used in teaching automation. In the modern era the application areas range from traditional process control to control of electric drives, applications in power electronics and electrical networks, smart grids, mobile networks and systems, digital systems, robotics, sensor networks, social networks etc. Especially communication in all levels is much more important than it used to be in traditional control engineering. The connection between signal processing and control engineering becomes even more important as it has usually been regarded. Similarly, information technology in general is very relevant in modern control applications. Consequently, there have been ideas to combine the two fields. For example, in Aalto University it is possible to apply to the bachelor level program Automation and Information technology. The studies are similar during the first year, after which the students will choose, whether they want to concentrate on information systems or automation.

In order to study, what has to be taught let us see the bachelor level curriculum (basic studies).

- Mathematics (25 cr.)
- Physics (10 cr.)
- Python-programming (5 cr.)
- C-programming (5 cr.)
- Hands-on course (8 cr.)
- Signals and systems (5 cr.)
- Mathematics software (2 cr.)
- Languages (5 cr.)
- Other (5 cr.)

The curriculum is more or less traditional with a few noticeable points. The amount of mathematics is considerable less than it used to be. Programming skills are considered important. Also, the Hands-on course is mandatory to all 1st year students. In that course the students work in groups with small introductory projects. For example, Arduino equipment and small robots are used to construct small devices. Electrical measurements and design, mechanical construction and design, programming and control are used in relatively simple settings. The idea is that the students get interested to appreciate the need to learn later more on the theoretical aspects, after first having done practical examples. (There remains a lot of unclear issues in the examples, e.g. in programming, signal processing and control, which should wake the students' interest to study more).

Still in bachelor level studies the major in automation and information technology looks like this.

- Bachelor seminar and thesis (10 cr.)
- Introduction to information technology (5 cr.)
- Introduction to Automation and systems technology (5 cr.)
- Automation 1 (5 cr.)
- Automation 2 (5 cr.)
- Laboratory course in automation (5 cr.)
- Control Engineering (5 cr.)
- Robotics (5 cr.)
- Basics of chemistry (5 cr.)
- Machine design (5 cr.)
- Electrical engineering and electronics (5 cr.)

Here the main issue to note is that both information technology and automation and systems technology are mandatory courses. After completing them the students will choose, which major of the two they will take.

The curriculum gets even more interesting in the major level. All teaching there is in English. In the Automation major the following seven courses are mandatory

- Embedded real-time systems (5 cr.)
- Product development (5 cr.)
- Project work course (10 cr.)
- Distributed and intelligent automation systems (5 cr.)
- Dynamical systems and identification (5 cr.)
- Stochastics and estimation (5 cr.)
- Advanced control (5 cr.)

In addition to this a broad selection of optional courses is available in a *course bank* from which the students can choose whatever courses they please to complete the major. There are certain *paths* which are suggested to take for certain specialization, but these are only suggestions. Like stated, the student is free to choose the courses.

The amount of product development oriented courses has been increased in the current program, and that is a clear plan in bringing the new teaching methods into play. Embedded real-time systems, product development and project work courses demonstrate this. Especially interesting is the project work course, which is a one-year-long course, where the students work in groups to carry out a projects given by research groups of the school. The problems are such that they are related to real research work, whenever this is possible (and usually it is). The instructors come from the research groups, and several sessions

and schooling to the instructors are given. The course includes lectures where project management, project planning, IPR rigts, group dynamics, business plan planning etc. are considered. The students choose a project manager from their group and there are project meetings on regular intervals (usually once in one or two weeks). Both agendas and minutes are written by the group members and that and other relevant material is managed in a suitable electronic project management system. In addition to that the course arranges mid-term presentations to the whole course where poster walks are arranged and each project is presented. A final gala is arranged in the end of the course. There the results are presented, including comments on how the project work felt and succeeded.

Peer evaluation is used as a part of determining the grade for each student. Here the students evaluate not only themselves but each other also. It is then the instructor who puts all information together and proposes the grades to the students. Usually that is the final grade, even though the responsible teacher of the course can intervene (usually not).

Although the project work course seems at a first glance a bit formal, it is actually a very good modern course. The students can use their imagination and talent to do new things, in a way that is supported by the methods and procedures of the work in start-ups and other companies. Some parts come near to the concept on *peer instruction* (Mazur, 1997), but as a whole the course goes even further in modern ideas of innovative, practical, yet theoretically challenging, engineering work.

In short, the new ideas implemented in the university level study curriculum in automation and control can be summarized as follows. A hands-on course is arranged already in the beginning of studies to give some practical work to the students and to wake their interest, motivation and need to study also theoretical courses. In theoretical courses more exercises and design problems are given, to be done alone or in a group in a time that is best suitable to the student(s). The amound of lectures is decreased and more weight in evaluation is given to the homeworks and design problems. Examination still exists, but it is more like a check that everybody in the courses masters the key material. In the examination there is not enough time to solve but rather simple exercises. It is much better to give time to the students, to work individually or in groups when they please. Usually, there is about two weeks time to solve one design problem.

The relatively large project work course was already explained above. It is a cornerstone in the new study program. It is not arranged only to the students of automatic control, but also to other students in different disciplines in the School of Electrical Engineering.

In addition to the courses where group work has a major role it is important to hear the voices of the "clients", that is industrial partners who will then hire the students for work. "Stakeholders' events" are arranged regularly to hear ideas and experiences on how the former students are doing and what aspects should be considered more in contemporary education. Similar idea is used by having contacts to the other universities, where automation is taught, in domestic level as well as abroad. Also, good contacts are formed to schools. That is a look into the future and is discussed in the next section.

4 Luma Activity

The future is not with us, it is in the young people. Even in the university level, the connections to schools are considered important. In Finland that is now organized in the Luma activity (Luma Centre Finland). "Luma" is actually very close to STEM education (Science, Technology, Engineering, Mathematics), the idea being to organize special courses and laboratory exercises to high school students and even to younger children. The purpose is to wake their interest in natural sciences and engineering, and later perhaps making the decision to start to study these disciplines. In the Luma centre Finland there are 13 different Luma units attached with universities across Finland. They all have different ways to operate: some have special Luma classes, where teachers can bring their students to do laboratory exercises; some arrange special courses and lectures to the students; some arrange student demonstrations and competitions e.g. in robotics etc. For example, in Luma centre Aalto there is a class (LUMARTS) with laboratory exercises in fields of biochemistry, chemistry, electrical engineering, meachanical engineering, and automation. Last year there were more than 2000 student visits in the class. Moreover, special courses on different topics are arranged regularly to high school students. Examples of such courses are micro- and nanosystems, robotics, space systems and mathematics.

In the LUMARTS class one class of problems is done with the Arduino platform, to get the students used to simple electronics applications and basic programming. Special courses in programming are also arranged, not only to the students but also to their teachers. The Luma activity is now pretty active in Finland and high interest and expectations are shown towards it, also in the political level.

To combine the ideas of the project work course described in the previous section and the Luma activity the following example can be given. In Figure 2 a laboratory example vessel has been presented. The exercise was constructed by a group of students in automation. The idea is to demonstrate the basic ideas of feedback control by using a liquid level control in a vessel as an example.

In the example the students can try tuning the liquid level manually, and they find it much more difficult than by an automatic PID-controller. It is clear that control mathematics (like the PID algorithm) cannot be taught at this level. However, the idea is again to wake an interest. Also, here the process represents something that also occurs in real process control and is then related to real work. (It is not an inverse double pendulum, which makes a nice demo, but is hardly related to most practical ap-

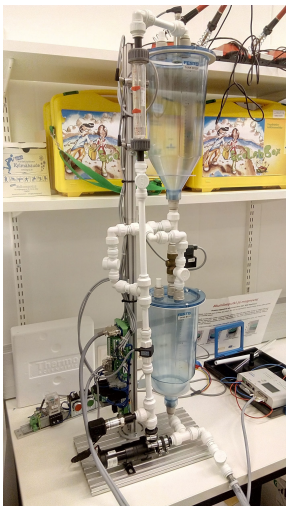


Figure 2. A simple process control system.

plications). Another example process constructed by the same project group is shown in Figure 3. It is a kind of a conveyor system, where a wooden ball is moved from one position to another by using a lift mechanism. The operation is controlled by a programmable logic controller, and one of the key targets in the exercise is to teach this kind of programming basics (ladder diagrams) to the students.

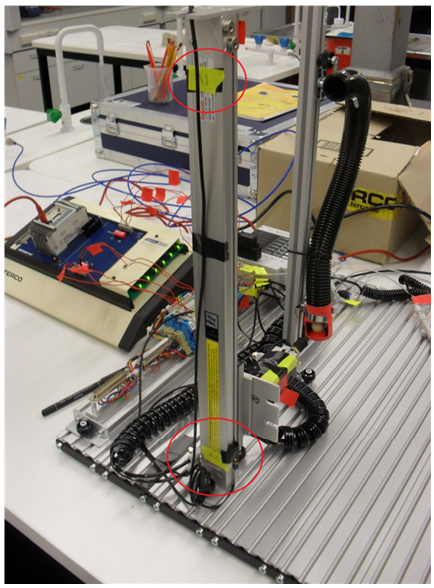


Figure 3. Autobygg system.

5 Conclusions

In the paper some modern concepts and ideas of control engineering education have been motivated and presented. The "new" methods are actually not very new and not very radical either. However, they are realistic and believed to meet the challenges that society sets to education in this field. It is also believed that the students in general like the more practical training they can get through the new kinds of courses. There is not yet enough student feedback to confirm or reject this hypothesis. Near future gives light to that issue also.

References

D. Ausubel. *Educational psychology: A cognitive view*. New York: Holt, Rinehart and Winston, 1968.

D. Boud and G. I. Feletti. *The Challenge of Problem-based Learning*. Psychology Press, 1977.

N. A. Kheir, K. J. Åström, D. Auslander, K. C. Cheok, G. F. Franklin, M. Masten, and M. Rabins. Control systems engineering education. *Automatica*, 32(2):147–166, 1996.

E. Mazur. *Peer Instruction: A User's Manual Series in Educational Innovation*. Prentice Hall, Upper Saddle River, NJ, 1997.

R. M. Murray. Future directions in control, dynamics and systems: Overview, grand challenges, and new courses. *European Journal of Control*, 9(2-3):144–158, 2003.

M. Pohjola, L. Eriksson, V. Hölttä, and T. Oksanen. Platform for monitoring and controlling educational laboratory exercises over internet. *In Proceedings of the 16th IFAC World Congress, Prague, Czech Republic*, 2005.

A. Rasila, M. Harjula, and K. Zenger. Automatic assessment of mathematics exercises: Experiences and future prospects. *In Reflektori 2007 Symposium on Engineering Education, Espoo, Finland*, pages 70–80, 2007.

V. Riihimäki, T. Ylöstalo, K. Zenger, and V. Maasalo. A new self-study course on the web: Basic mathematics in control. *In Proceedings of the IFAC Symposium ACE 2003 (J. Lindfors, Ed.), Oulu, Finland, 2003*, pages 149–153, 2003.

K. Zenger. Control engineering, system theory and mathematics: the teacher's challenge. *Journal of Engineering Education*, 32(6):687–694, 2007.

A Simplified Model of an Activated Sludge Process with a Plug-Flow Reactor

Jesús Zambrano[1] Bengt Carlsson[2] Stefan Diehl[3] Emma Nehrenheim[1]

[1]School of Business, Society and Engineering, Mälardalen University, Västerås, Sweden,
jesus.zambrano@mdh.se
[2]Department of Information Technology, Uppsala University, Uppsala, Sweden.
[3]Centre for Mathematical Sciences, Lund University, Lund, Sweden.

Abstract

The analysis of a simplified activated sludge process (ASP) with one main dissolved substrate and one main particulate biomass has been conducted in steady-state conditions. The ASP is formed by a plug-flow reactor and a settler tank. The biomass growth rate is described by a Monod function. For this process, it is not possible to get an explicit expression for the effluent substrate concentration when the process is subject to a fixed sludge age. However, when the substrate concentration of the influent is much greater than that of the effluent, an approximate and explicit relation between them is obtained. Numerical examples with two models for the settler are presented. One model is the ideal settler, which assumes a complete thickening of the sludge. The other model includes hindered settling and sludge compression. Numerical results show the effectiveness and the limitations of the proposed solution under these scenarios.

Keywords: bioreactor, clarifier, sludge blanket, sludge age

1 Introduction

Steady-state modeling and analysis of ASPs have been extensively studied during the last 50 years. One important reason is that the steady-state analysis of a dynamic model can provide initial values for process operation and optimization.

Generally, the mixing regime in an ASP reactor neither behaves as a completely stirred tank reactor (CSTR) nor as a plug-flow reactor (PFR), but in some sense in between (Tsai and Chen, 2013). In a CSTR, the reactor content is well stirred, so it is assumed that the concentration in the effluent is the same as in the reactor. In a PFR, the key assumption is that the fluid is perfectly mixed in the radial direction and in the axial direction only the transportation of the fluid is considered. Therefore, a PFR can be seen as a series of infinitely thin CSTRs, each with a uniform and different composition than the neighbouring one (Schmidt, 1998). It is expected that a PFR with a volume smaller than several CSTRs in series will give the same performance (Zambrano et al., 2015).

Compared to the classical ASP configuration, i.e. ASP with one CSTR, an explicit (steady-state) solution for an ASP formed by a PFR seems to not be possible to obtain (Diehl et al., 2017). However, some attempts have been made in the analysis of this process. For example, some implicit and approximate expressions for the effluent substrate were presented by San (1989), where the expressions were compared with numerical solutions. Computer techniques to solve the problem of a PFR in an ASP when considering the PFR as a large number of bioreactors in series are shown by Muslu (2000). Design graphs and numerical examples were presented as guidelines to size the process. On the other hand, a study of an ASP formed by a PFR could be seen as an approximation of an ASP with several CSTRs in series (Erickson and Fan, 1968; Zambrano and Carlsson, 2014).

A study of the relationship between the influent and effluent of an ASP formed by one and two CSTRs in series was presented in Zambrano and Carlsson (2014). The study mentions that it does not seem possible to find explicit solutions for the effluent substrate concentration for two or more bioreactors. That work was the motivation that led to the development of Diehl et al. (2017) and the current study.

A steady-state analysis of an ASP formed by using a PFR and a settler was recently studied in Diehl et al. (2017). The study considers and compares two different settler models. One is the ideal settler, which assumes an unlimited flux capacity, i.e. the settler is always considered to be overdimensioned. The other model, recently published by Diehl et al. (2016), here referred to as DZC settler model, includes hindered and compressive settling, which means that a limited flux capacity is modelled. Both numerical and, in some cases, analytical results are obtained. A comparison with an ASP formed by a single CSTR is also shown by Diehl et al. (2017).

In the present work, we continue the analysis of an ASP consisting of a PFR and a settler. For the ideal settler case, the steady-state solution is presented with explicit approximate formulas when the influent substrate concentration is much greater than the effluent substrate concentration. Under the same assumption, we also present an explicit formula for the effluent substrate concentration as a function of the influent substrate concentration when the

sludge age is fixed. This formula can be used for both settler models.

Nomenclature

A	vertical cross-sectional of PFR $[\mathrm{m}^2]$
A_S	horizontal cross-sectional of settler $[\mathrm{m}^2]$
B	depth of thickening zone $[\mathrm{m}]$
H	height of clarification zone $[\mathrm{m}]$
K_s	half-saturation constant $[\mathrm{kg/m}^3]$
Q	influent volumetric flow rate $[\mathrm{m}^3/\mathrm{s}]$
S	dissolved substrate concentration $[\mathrm{kg/m}^3]$
U_{z_sb}	function defined in (12) $[\mathrm{kg/m}^3]$
V_R	volume of PFR $[\mathrm{m}^3]$
X	particulate biomass concentration $[\mathrm{kg/m}^3]$
$X^\infty_{z_\mathrm{sb}}$	parameter in (12) $[\mathrm{kg/m}^3]$
Y	yield constant $[-]$
h	length of PFR $[\mathrm{m}]$
q	bulk velocity in the thickening zone $[\mathrm{m/s}]$
$\hat{q}_{z_\mathrm{sb}}$	parameter in (12) $[\mathrm{m/s}]$
$\check{q}_{z_\mathrm{sb}}$	parameter in (12) $[\mathrm{m/s}]$
r	recycle ratio $[-]$
w	wastage ratio $[-]$
x	horizontal distance from feed in PFR $[\mathrm{m}]$
z	depth from feed level in settler $[\mathrm{m}]$
Greek letters	
μ	Monod function $[1/\mathrm{s}]$
μ_max	maximum specific growth rate $[1/\mathrm{s}]$
θ	sludge age $[\mathrm{s}]$
Subscripts	
$\square_0$	defined constant value
$\square_\mathrm{e}$	effluent
$\square_\mathrm{in}$	influent
$\square_\mathrm{r}$	recycle
$\square_\mathrm{sb}$	sludge blanket
Superscripts	
$\square^*$	PFR steady-state concentration
$\overline{\square}$	PFR influent concentration

The paper is organized as follows. A description of the ASP with a PFR and a settler is presented in Section 2, including the steady-state mass balances and the definitions for the ideal and DZC settler models. In Sections 3 and 4, we review from Diehl et al. (2017) the equations describing the steady-state conditions of the ASP for both settler models. Section 5 contains an approximate explicit expression for the effluent substrate concentration. Numerical examples are shown in Section 6 and conclusions are drawn in Section 7.

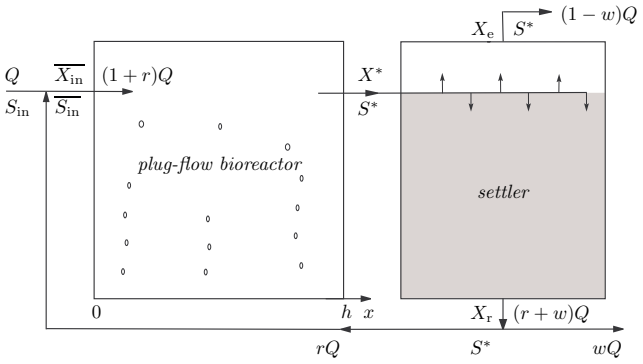


Figure 1. The activated sludge process consisting of a PFR and a settler. The steady-state variables are shown as well as the horizontal x-axis of the PFR.

2 The Activated Sludge Process

For the ASP we consider using a PFR coupled with a settler, see Figure 1, where the recycling flows to the reactor. The PFR has a constant vertical cross-sectional area A and length h, so the volume is $V_\mathrm{R} = Ah$. The variable x is used to denote the horizontal axis in the PFR from the inlet ($x = 0$) to the outlet ($x = h$). Where the concentrations at location x can be denoted as $S(x)$ and $X(x)$ in the PFR.

We assume two constituents, namely one particulate biomass X and one dissolved substrate S. The influent volumetric flow rate and substrate concentration are denoted by Q and S_in, respectively. It is assumed that no biomass is present in the influent ($X_\mathrm{in} = 0$). The PFR input concentrations are denoted by $\overline{S_\mathrm{in}}$ and $\overline{X_\mathrm{in}}$, and the PFR outputs by S^* and X^*. It is assumed that no reactions are taking place in the settler, so that only particulate biomass is influenced. The substrate concentration is thus unchanged and therefore equal to S^* throughout the settler. The effluent at the top of the settler is X_e and the recycle concentration is X_r. The recycle flow rate is rQ and the waste flow rate is wQ, where $r > 0$ and $0 < w \leq 1$. The kinetics in the PFR are described by using the Monod function (Monod, 1949)

$$\mu(S) = \mu_\mathrm{max} \frac{S}{K_\mathrm{s} + S}, \tag{1}$$

where μ_max is the maximum specific growth rate and K_s is the half-saturation constant. It is assumed that the biomass death is negligible.

The sludge age θ of the process is defined as the amount of biomass in the bioreactor divided by the removed biomass per unit time, and is expressed as

$$\theta = \frac{A \int_0^h X(x)dx}{wQX_\mathrm{r}}. \tag{2}$$

2.1 Mass balances and expression for the sludge age

The three mass balances of the process in steady state with $X_e = 0$ are

$$Q(1+r)\overline{S_{in}} = QS_{in} + rQS^*, \quad (3)$$
$$Q(1+r)\overline{X_{in}} = rQX_r, \quad (4)$$
$$Q(1+r)X^* = (r+w)QX_r. \quad (5)$$

Applying the conservation of mass in the PFR we get

$$\frac{Q(1+r)}{A}\frac{dS}{dx} = -\mu[S(x)]\frac{X(x)}{Y}, \quad (6)$$
$$\frac{Q(1+r)}{A}\frac{dX}{dx} = \mu[S(x)]X(x), \quad (7)$$

where Y refers to the yield constant. The following boundary conditions hold: $X(0) = \overline{X_{in}}$, $S(0) = \overline{S_{in}}$, $X(h) = X^*$ and $S(h) = S^*$. Combining Equations (6) and (7) together with the boundary conditions, one gets

$$\frac{Q(1+r)}{A}\frac{d(YS+X)}{dx} = 0$$
$$\implies \quad Y\overline{S_{in}} + \overline{X_{in}} = YS(x) + X(x) = YS^* + X^*. \quad (8)$$

By solving for $X(x)$ in Equation (8) and substituting it into (6), using $V_R = Ah$ and integrating, we get the following equation for the PFR:

$$-Q(1+r)Y\int_{\overline{S_{in}}}^{S^*}\frac{d\sigma}{\mu(\sigma)\left[\overline{X_{in}} + Y(\overline{S_{in}} - \sigma)\right]} = V_R$$
$$\iff \quad f(S^*, r, w) = V_R, \quad (9)$$

where

$$f(S^*, r, w) = \frac{Q(1+r)}{\mu_{max}}\left[P\ln\left(\frac{a(S_{in}+rS^*)}{S^*(1+r)}\right) + \ln(a)\right], \quad (10)$$

$$P = P(S^*, r, w) = \frac{K_s w(1+r)}{S_{in}(r+w) - S^* r(1-w)},$$
$$a = a(r, w) = \frac{r+w}{r}.$$

We can obtain the sludge age by rewriting the integral in Equation (2) using Equation (6). Then we have (see Diehl et al. (2017))

$$\theta = \frac{A}{wQX_r}\left[\frac{-Q(1+r)Y}{A}\int_{\overline{S_{in}}}^{S^*}\frac{(K_s+\sigma)d\sigma}{\mu_{max}\sigma}\right]$$
$$= \frac{1}{\mu_{max}}\left[1 + \frac{(1+r)K_s}{(S_{in}-S^*)}\ln\left(\frac{S_{in}+rS^*}{S^*(1+r)}\right)\right]. \quad (11)$$

2.2 The settler

Ideal settler model. For an ideal settler we assume that all the sludge fed to the settler will always pass through the thickening zone, regardless of the amount of incoming sludge and the recycle and waste flows. Although in many cases unrealistic, this model could work well when the settler is over-sized.

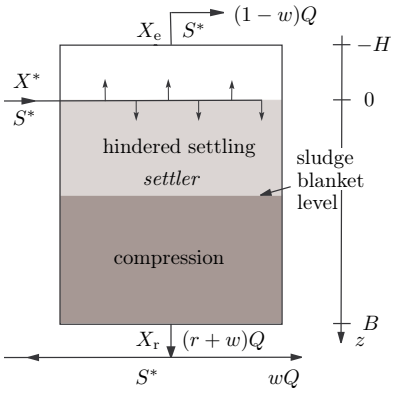


Figure 2. The DZC settler. The steady-state variables are shown as well as the vertical z-axis of the settler.

DZC settler model. The processes in the settler are described by a steady-state approximation of a partial differential equation (PDE) which includes a hindered settling velocity function and a compression function (Bürger et al., 2011). The behavior of a real settler can be divided into three qualitatively different operations: underloaded, overloaded and normal operation. By normal operation we mean that all the biomass fed to the settler is conveyed through the thickening zone and that there exists a sludge blanket in the thickening zone, see Figure 2. In this work we only study the steady-state solutions under normal operation and therefore set $X_e = 0$.

The following simple relationship is a reasonable approximation obtained from the steady-state solutions that have a sludge blanket in the thickening zone (Diehl et al., 2016)

$$X_r = U_{z_{sb}}(q) := X^{\infty}_{z_{sb}}\left(1 + \frac{\hat{q}_{z_{sb}}}{q + \check{q}_{z_{sb}}}\right), \quad (12)$$

where q is the bulk velocity in the thickening zone, defined as

$$q = q(r, w, Q, A_S) := \frac{Q(r+w)}{A_S}, \quad (13)$$

where $X^{\infty}_{z_{sb}}$, $\hat{q}_{z_{sb}}$ and $\check{q}_{z_{sb}}$ are parameters which depend on the chosen sludge blanket level z_{sb}. A_S is the settler constant horizontal cross-sectional area, see model details in Diehl et al. (2016).

3 ASP with ideal settler model

3.1 Steady-state solutions

From the mass balances of the process (5), (8) and (9), the steady-state equations for an ASP with ideal settler can be expressed as (ignoring the variables $\overline{S_{in}}$ and $\overline{X_{in}}$; these can

be obtained from (3) and (4))

$$S^* = S_{\text{in}} - \frac{w}{Y} X_{\text{r}}, \tag{14}$$

$$X^* = \frac{r+w}{1+r} X_{\text{r}}, \tag{15}$$

$$V_{\text{R}} = f(S^*, r, w), \tag{16}$$

where $f(S^*, r, w)$ is given by (10). Equation (16) is solved for $S^* = S^*(r,w)$, then Equation (14) gives $X_{\text{r}} = X_{\text{r}}(r,w)$ and Equation (15) gives $X^* = X^*(r,w)$. Note that all these variables are two-parameter solutions of the control variables r, w. Note also from Equations (9) and (10) that S^* is expressed implicitly. If $S_{\text{in}} \gg S^*$ is assumed, we have the following results.

Theorem 1. *Given an ASP with an ideal settler described by Equations* (14)–(16)*. If $S_{\text{in}} \gg S^*$ then the solution of Equations* (14)–(16) *can be expressed explicitly as*

$$S^* = S^*(r,w) = \frac{(r+w)S_{\text{in}}}{r\left[(1+r)\exp(\beta) - (r+w)\right]}, \tag{17}$$

$$X_{\text{r}} = X_{\text{r}}(r,w) = \frac{Y}{w}\left(S_{\text{in}} - S^*(r,w)\right), \tag{18}$$

$$X^* = X^*(r,w) = \frac{(r+w)Y}{(1+r)w}\left(S_{\text{in}} - S^*(r,w)\right). \tag{19}$$

where

$$\beta = \frac{S_{\text{in}}(r+w)}{K_{\text{s}} w(1+r)}\left[\frac{V_{\text{R}}\mu_{\max}}{Q(1+r)} - \ln\left(\frac{r+w}{r}\right)\right],$$

and when the denominator in Equation (17) *is positive.*

Proof. The assumption implies that Equation (16) can be expressed as (cf. Equations (9) and (10))

$$\frac{Q(1+r)}{\mu_{\max}}\left[\frac{1}{G}\ln\left(\frac{a(S_{\text{in}} + rS^*)}{S^*(1+r)}\right) + \ln(a)\right] = V_{\text{R}}, \tag{20}$$

where

$$G = \frac{S_{\text{in}}(r+w)}{K_{\text{s}} w(1+r)} \quad \text{and} \quad a = \frac{r+w}{r}.$$

Solving (20) for S^* we get

$$G\left(\frac{V_{\text{R}}\mu_{\max}}{Q(1+r)} - \ln(a)\right) = \ln\left(\frac{a(S_{\text{in}} + rS^*)}{S^*(1+r)}\right).$$

For simplicity we set $\beta = G\left(\frac{V_{\text{R}}\mu_{\max}}{Q(1+r)} - \ln(a)\right)$, then we have

$$\begin{aligned}
\exp(\beta) = \frac{a(S_{\text{in}} + rS^*)}{S^*(1+r)} \quad &\Longleftrightarrow \\
S^*(1+r)\exp(\beta) = a(S_{\text{in}} + rS^*) \quad &\Longleftrightarrow \\
S^* = \frac{aS_{\text{in}}}{(1+r)\exp(\beta) - ar} \quad &\Longleftrightarrow \\
S^* = \frac{(r+w)S_{\text{in}}}{r\left[(1+r)\exp(\beta) - (r+w)\right]},&
\end{aligned}$$

if the denominator is positive.

Once S^* is obtained, X_{r} and X^* are given from Equations (14) and (15), respectively. □

3.2 Substrate input-output relationship for constant sludge age

The two-parameter solution of Equations (14)–(16) (or (17)–(19) in Theorem 1) means that two additional equations can be imposed to define the operation conditions. We are interested in investigating the steady-state solutions of the process for different values of S_{in} for a constant sludge age θ_0. For S_{in} as a variable, we have six variables to take into consideration: $S^*, X^*, X_{\text{r}}, r, w$, and S_{in}. However, to get a one-parameter solution with S_{in} as a parameter, we can add the following to Equations (14)–(16):

$$r = r_0, \tag{21}$$

$$\frac{1}{\mu_{\max}}\left[1 + \frac{(1+r)K_{\text{s}}}{(S_{\text{in}} - S^*)}\ln\left(\frac{S_{\text{in}} + rS^*}{S^*(1+r)}\right)\right] = \theta_0. \tag{22}$$

Since $r = r_0$ is constant, Equation (22) defines implicitly $S^* = S^*(S_{\text{in}})$, then Equation (16) gives $w = w(S_{\text{in}})$, Equation (14) gives $X_{\text{r}} = X_{\text{r}}(S_{\text{in}})$ and Equation (15) gives $X^* = X^*(S_{\text{in}})$.

4 ASP with DZC settler model

4.1 Steady-state solutions

The mass balances of the system considering the DZC settler model have to include Equation (12) in order to get a sludge blanket in the thickening zone. The steady-state equations are then expressed as

$$S^* = S_{\text{in}} - \frac{w}{Y} U_{z_{\text{sb}}}(q), \tag{23}$$

$$X^* = \frac{r+w}{1+r} U_{z_{\text{sb}}}(q), \tag{24}$$

$$X_{\text{r}} = U_{z_{\text{sb}}}(q), \tag{25}$$

$$V_{\text{R}} = f(S^*, r, w). \tag{26}$$

Straightforward calculations give that the expression for the sludge age is the same as for the ideal settler model (cf. Equation (11) and Diehl et al. (2017)).

4.2 Substrate input-output relation for constant sludge age

As in the case of an ASP with ideal settler, we are interested in the solution of the process for a constant sludge age for different values of S_{in}. By imposing $\theta(r, S_{\text{in}}, S^*) = \theta_0$ we get a one-parameter solution. Note that we cannot impose another equation (e.g. $r = r_0$) as we did for the ideal settler model, since we have Equation (12) controlling the sludge blanket.

Hence, Equations (22), (23) and (26) are solved for $S^* = S^*(S_{\text{in}})$, $r = r(S_{\text{in}})$ and $w = w(S_{\text{in}})$. Then, Equation (24) gives $X^* = X^*(S_{\text{in}})$ and Equation (25) gives $X_{\text{r}} = X_{\text{r}}(S_{\text{in}})$.

5 An approximation for S^* given θ_0

Note that S^* is given implicitly in Equation (22), and will depend on S_{in}, r and θ_0. This equation can, however, be solved explicitly for S^* if we make an assumption.

Theorem 2. *Given an ASP described by Equations* (14)–(16) *(for an ideal settler) or by* (23)–(26) *(for a DZC settler), and where θ is given by* (11)*. Assume that θ is fixed to θ_0, i.e. Equation* (22) *is imposed. If $S_{\text{in}} \gg S^*$, then the following simple expression for S^* holds:*

$$S^* = \frac{S_{\text{in}}}{(1+r)\exp(\alpha S_{\text{in}}) - r}, \tag{27}$$

where

$$\alpha = \frac{\theta_0 \mu_{\max} - 1}{K_{\text{s}}(1+r)}.$$

Proof. Assuming $S_{\text{in}} \gg S^*$, Equation (22) can be written as

$$\frac{1}{\mu_{\max}}\left[1 + \frac{(1+r)K_{\text{s}}}{S_{\text{in}}} \ln\left(\frac{S_{\text{in}} + rS^*}{S^*(1+r)}\right)\right] = \theta_0,$$

solving for S^* gives

$$S^* = \frac{S_{\text{in}}}{(1+r)\exp(\alpha S_{\text{in}}) - r},$$

where $\alpha = (\theta_0 \mu_{\max} - 1)/(K_{\text{s}}(1+r))$. □

6 Numerical example

We assume that the ASP has the following constants and parameters: $V_{\text{R}} = 3000\,\text{m}^3$, $Q = 1000\,\text{m}^3/\text{h}$, $\mu_{\max} = 0.17\,\text{h}^{-1}$, $K_{\text{s}} = 0.05\,\text{kg/m}^3$, $Y = 0.7$. For the DZC settler model we let: $B = 3\,\text{m}$, $z_{\text{sb}} = 1\,\text{m}$, $X_1^{\infty} = 6.52\,\text{kg/m}^3$, $\hat{q}_1 = 0.32\,\text{m/h}$, $\check{q}_1 = 0.45\,\text{m/h}$. The latter constants were obtained with standard parameters for the hindered settling and compression functions and the procedure in Diehl et al. (2016).

Numerical solutions of the model equations will now be compared with the approximate solutions given by Theorems 1 and 2. The numerical solutions are obtained with `fsolve`, a function in Matlab which solves systems of non-linear equations.

6.1 Theorem 1

This case deals with an ASP with an ideal settler model. Figure 3 shows the numerical (without approximation, i.e. without the assumption $S_{\text{in}} \gg S^*$) and approximated solutions given by Theorem 1 for S^*, X^* and X_{r}. That is, we compare the results from Equations (14)–(16) with results from Equations (17)–(19). The influent substrate concentration is set to $S_{\text{in}} = 0.1\,\text{kg/m}^3$. The results are shown for an interval of values of r and for some values of the wastage ratio w.

Note that in plot (a), for higher values of S^*, the difference between the values given by Theorem 1 and the values from the solution with no approximation becomes larger. The same effect can be seen in plot (b) for X^* and in plot (c) for X_{r}. S^* starts to decrease for higher values in r. Then, the difference between values with and without approximation starts to decrease.

6.2 Theorem 2

For the ideal settler model, Figure 4 shows numerical and approximated solutions given by Theorem 2. Equation (22) is solved for $S^* = S^*(S_{\text{in}})$ at an interval of values for S_{in}. The recirculation is set to $r = r_0 = 1$ (cf. Equation (21)), and we set $\theta_0 = 16\,\text{h}$. Note that, for higher values of S_{in}, the values for S^* given by Theorem 2 are closer to those given by the solution of the process with no approximation.

For the DZC settler model, Equations (22), (23) and (26) are solved for $S^* = S^*(S_{\text{in}}), r = r(S_{\text{in}})$ and $w = w(S_{\text{in}})$ at an interval of values for S_{in}. Figure 5 shows the numerical and approximated solutions for $S^* = S^*(S_{\text{in}})$. We set $\theta_0 = 6.5\,\text{h}$ and show some results for some values of the settler area A_{S}.

Note that, for a given S_{in} and when a higher A_{S} is used, the solution with no approximation gives a higher recycle concentration X_{r} (see Equations (12) and (13)). This means that we have a more thickened sludge, which gives a lower S^* (see Equation (23)). Therefore, S^* becomes much lower compared to S_{in} as A_{S} increases. Hence, the values from Theorem 2 are much closer to the solution of the model equations without approximation.

7 Conclusions

An ASP formed by a PFR and a settler has been studied in steady-state operation. It is shown that explicit, but approximate, solutions can be obtained for the case of an ideal settler under the assumption that the influent substrate concentration is much greater than the effluent one. With this assumption it is also possible to obtain an explicit expression for the effluent concentration as a function of the influent one under the operating condition that the sludge age should be maintained at a specific value. Numerical examples show the performances of the simpler explicit expressions under two different models for the settler and hence when the simpler formulas can be used. Further research might be focused on considering the decayed particulate biomass as an additional constituent in the process.

Acknowledgment

The research leading to these results has received funding from the Knowledge Foundation (20140168) under the

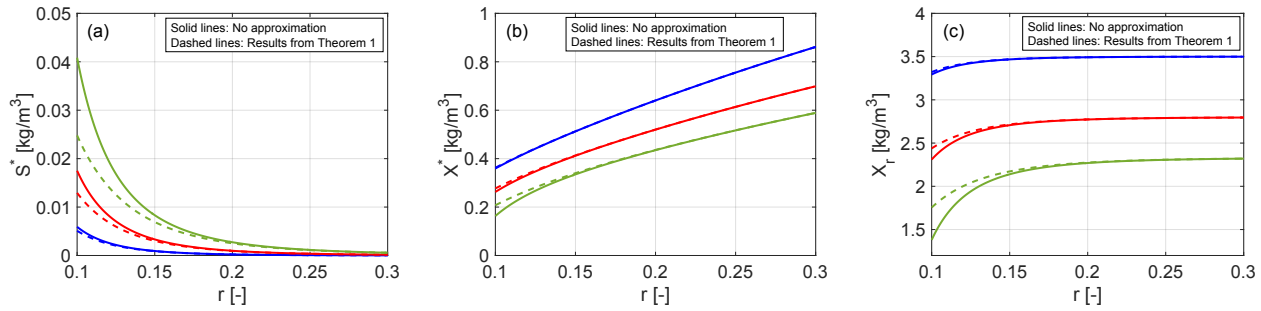


Figure 3. PFR with ideal settler model. Comparison between the numerical (no approximation) and the approximated solutions given by Theorem 1 as functions of r. The results are shown for three values of the wastage: $w = 0.02$ (in blue), $w = 0.025$ (in red), $w = 0.03$ (in green). The influent substrate concentration is fixed to $S_{\text{in}} = 0.1$ kg/m^3.

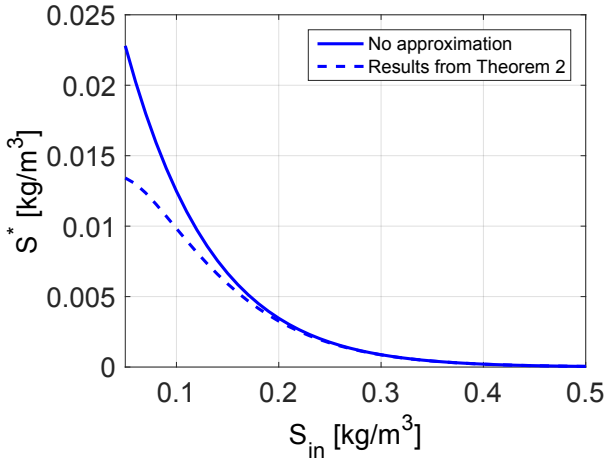


Figure 4. PFR with ideal settler model. Comparison between the numerical and the approximated solutions given by Theorem 2 as functions of S_{in}. The recirculation is fixed to $r_0 = 1$ and the sludge age is kept to $\theta_0 = 16$ h.

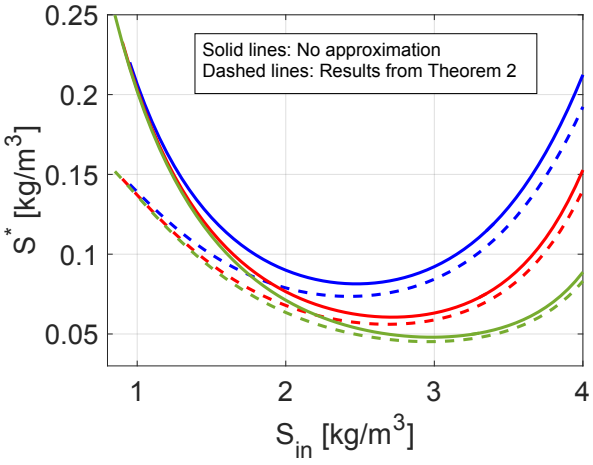


Figure 5. PFR with DZC settler model. Comparison between the numerical (no approximation) and the approximated solutions given by Theorem 2 as functions of S_{in} for some values of the settler area A_S [m^2]: 500 m^2 (in blue), 1500 m^2 (in red), 3000 m^2 (in green). For every curve, the sludge age is kept to $\theta_0 = 6.5$ h.

project More Sense, ABB, Mälarenergi, Flexiclean and Structor Miljöteknik.

References

R. Bürger, S. Diehl, and I. Nopens. A consistent modelling methodology for secondary settling tanks in wastewater treatment. *Water Res.*, 45(6):2247–2260, 2011. doi:10.1016/j.watres.2011.01.020.

S. Diehl, J. Zambrano, and B. Carlsson. Steady-state analysis of activated sludge processes with a settler model including sludge compression. *Water Research*, 88:104–116, 2016. doi:10.1016/j.watres.2015.09.052.

S. Diehl, J. Zambrano, and B. Carlsson. Steady-state analyses of activated sludge processes with plug-flow reactor. *Journal of Environmental Chemical Engineering*, 5(1):795–809, 2017. doi:10.1016/j.jece.2016.06.038.

L. E. Erickson and L.-t. Fan. Optimization of the hydraulic regime of activated sludge systems. *Journal Water Pollution Control Federation*, 40(3):345–362, 1968. ISSN 00431303.

J. Monod. The growth of bacterial cultures. *Annu. Rev. Microbiol.*, 3(1):371–394, 1949.

Y. Muslu. Numerical approach to plug-flow activated sludge reactor kinetics. *Comput. Biol. Med.*, 30:207–223, 2000. doi:10.1016/S0010-4825(00)00009-3.

H. Ali San. A kinetic model for ideal plug-flow reactors. *Water Res.*, 23(5):647–654, 1989. doi:10.1016/0043-1354(89)90031-6.

L. Schmidt. *The Engineering of Chemical Reactions*. Oxford University Press, 1998.

D. D. W. Tsai and P. H. Chen. Differentiation criteria study for continuous stirred tank reactor and plug flow reactor. *Theoretical Foundations of Chemical Engineering*, 47(6):750–757, 2013. doi:10.1134/s0040579513060122.

J. Zambrano and B. Carlsson. Steady-state analysis of simple activated sludge processes with Monod and Contois growth

kinetics. In *IWA Special International Conference: "Activated Sludge - 100 Years and Counting"*, Essen, Germany, 2014.

J. Zambrano, B. Carlsson, and S. Diehl. Optimal steady-state design of zone volumes of bioreactors with Monod growth kinetics. *Biochem. Eng. J.*, 100:59–66, 2015. doi:10.1016/j.bej.2015.04.002.

Monitoring a Secondary Settler using Gaussian Mixture Models

Jesús Zambrano[1] Oscar Samuelsson[2,3] Bengt Carlsson[3]

[1]School of Business, Society and Engineering, Mälardalen University, Box 883, 72123 Västerås, Sweden,
jesus.zambrano@mdh.se
[2]IVL Swedish Environmental Research Institute, P.O. Box 210 60, 10031 Stockholm, Sweden.
[3]Department of Information Technology, Uppsala University, Box 337, 75105 Uppsala, Sweden.

Abstract

This paper presents a method for monitoring the sludge profiles of a secondary settler using a Gaussian Mixture Model (GMM). A GMM is a parametric probability density function represented as a weighted sum of Gaussian components densities. To illustrate this method, the current approach is applied using real data from a sensor measuring the sludge concentration as a function of the settler level at a wastewater treatment plant (WWTP) in Bromma, Sweden. Results suggest that the GMM approach is a feasible method for monitoring and detecting possible disturbances of the process and fault situations such as sensor clogging. This approach can be a valuable tool for monitoring processes with a repetitive profile.

Keywords: signal monitoring, fault detection, clarifier, sludge profile

1 Introduction

The effluent water quality and efficient operation of resources are important aspects considered in the operation of a wastewater treatment plant (WWTP). Process monitoring and detection of abnormal conditions are crucial tasks, since they can help to improve the process performance (Olsson et al., 2014).

The sedimentation is an important process that determines the performance of the activated sludge process (ASP). The sedimentation is given by a secondary settler tank (SST), also called clarifier, which use gravity to separate the sludge (biomass) component from the treated water (liquid). Different approaches for predicting the SST behavior includes one, two or three-dimensional dynamic models. However, the prediction of the concentration profiles is still far from satisfactory (Li and Stenstrom, 2014), which makes the SST monitoring a complex task. Some examples of methods applied to monitor SSTs include image analysis (Grijspeerdt and Verstraete, 1997) and model-based approaches (Traoré et al., 2006; Yoo et al., 2002).

In the last two decades, a research field called *Machine Learning* has gained especial attention. The main scope with Machine Learning is to develop methods that can automatically detect patterns in data (learning), and then to use the uncovered patters to predict future data (Murphy, 2012). There are many different approaches in machine learning including decision trees, data clustering, neural networks, Gaussian process regression, Gaussian mixture models.

The authors proposed in Zambrano et al. (2015) an approach for monitoring a SST using Gaussian Process Regression (GPR), giving useful information about the status of the settler. GPR is a non-parametric regression method where data prediction is given as a probability density function. Hence, the predicted value comes with a variance estimate, interpreted as an uncertainty of the prediction. The method is thoroughly described by, for example, Rasmussen and Williams (2005) and Murphy (2012), and has gained large interest within the machine learning community for applications such as fault detection of environmental signals (Osborne et al., 2012), signal prediction (Grbić et al., 2013a,b) and control of bioreactors (Kocijan and Hvala, 2013).

In this work, we propose an alternative method for monitoring the process presented by Zambrano et al. (2015) based on a Gaussian Mixture Model (GMM). GMM is a parametric probability model for density estimation using a mixture of Gaussian distributions (Bishop, 2007). In this way, the GMM can describe a set of data using the combination of Gaussian distributions. Diverse applications of GMM can also be found in literature, for example in sensor monitoring (Zhu et al., 2014), fault detection and diagnosis (Yu, 2012).

The paper is organized as follows. First, a general introduction to GMM is presented, including a fault detection criteria based on the GMM formulation. Then, the problem of monitoring a secondary settler is presented as case study. Next, results and discussions are presented. Finally, some conclusions are drawn.

2 Materials and Methods

This section first presents the basics of Gaussian Mixture Models (GMM). Further, a GMM-based fault detection criteria is defined.

2.1 Gaussian Mixture Models

Assume we have a data vector x with N independent observations from a certain process. In a GMM, the total distribution of data is modeled as a sum (or mixture) of several Gaussian distributions with mean μ_k and covariance matrix σ_k. Hence, the model can be expressed as

(Murphy, 2012)

$$p(\mathrm{x}_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathrm{x}_i|\mu_k, \sigma_k),\ i = 1, ..., N \tag{1}$$

where each Gaussian distribution is denoted by $\mathcal{N}(\mathrm{x}_i|\mu_k, \sigma_k)$. The expression (1) is a combination of K Gaussian distributions, since we are taking a weighted sum. The mixing weights π_k must satisfy $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. The resulting function $p(\mathrm{x}_i)$ is a probability density function (pdf) from observing the data x_i.

When the value of K-groups is specified, the GMM parameters π_k, μ_k and σ_k can be inferred by using the iterative Expectation-Maximization (EM) algorithm applied to Gaussian Mixtures (Murphy, 2012), which can be summarized in Algorithm 1.

Algorithm 1 EM for Gaussian mixtures

1: Initialize $\mu_k^1, \sigma_k^1, \pi_k^1$ and set $i = 1$.
2: **while** not converged **do**
3: Compute $\gamma(z_{nk})$. ▷ Expectation step
4: Compute $\mu_k^{i+1}; \pi_k^{i+1}; N_k; \sigma_k^{i+1}$. ▷ Maximization step
5: $i \leftarrow i + 1$.
6: **end while**

The expressions used in Algorithm 1 are

$$\gamma(z_{nk}) = \frac{\pi_k^i \mathcal{N}(\mathrm{x}_n|\mu_k^i, \sigma_k^i)}{\sum_{j=1}^{K} \pi_j^i \mathcal{N}(\mathrm{x}_n|\mu_j^i, \sigma_j^i)}, n = 1, ..., N; k = 1, ..., K \tag{2}$$

$$\mu_k^{i+1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathrm{x}_n, \tag{3}$$

$$\pi_k^{i+1} = \frac{N_k}{N}, \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk}), \tag{4}$$

$$\sigma_k^{i+1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathrm{x}_n - \mu_k^{i+1}\right) \left(\mathrm{x}_n - \mu_k^{i+1}\right)^T. \tag{5}$$

One way to assign a value for K is using the silhouette criterion, see details in Rousseeuw (1987). The silhouette value S estimates how similar samples are in one cluster to samples in another cluster. S ranges from -1 (data misclassified) to $+1$ (data well-clustered), where S close to zero means that the clusters are indistinguishable.

2.2 GMM based fault detection criteria

When implementing a GMM to a group of data, the main idea is to compute a residual r so to monitor and decide between normal and abnormal profiles in the process. We assume that r belongs to one out of two different hypothesis: H_0 and H_1. Hence, the problem can be expressed by the classical binary hypothesis testing problem

$$\begin{aligned} H_0 &: r \leq h \\ H_1 &: r > h \end{aligned} \tag{6}$$

where H_0 refers to the non-faulty (normal) condition hypothesis, H_1 refers to the faulty (abnormal) condition hypothesis, and h is a predefined threshold. The aim is to decide if the system has changed between H_0 and H_1 when changes in the dynamic of the process are presented. It is assumed that H_0 and H_1 are equally likely.

For monitoring a group of profile data, each of them with N samples, we propose a GMM based residual r as detailed in Algorithm 2.

Algorithm 2 GMM-based residual calculation

1: Collect a group of M-profiles in non-faulty conditions.
2: Set K and compute the iterative EM algorithm (see Algorithm 1) to get π_k, μ_k, σ_k.
3: **while** monitoring a new profile **do**
4: **for** every profile **do**
5:

$$r = \frac{1}{p(\mathrm{x}; \pi_{1:K}, \mu_{1:K}, \sigma_{1:K})}, \tag{7}$$

where

$$p(\mathrm{x}; \pi_{1:K}, \mu_{1:K}, \sigma_{1:K}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathrm{x}_n|\mu_k, \sigma_k). \tag{8}$$

6: **end for**
7: **end while**

As given by expression (6), a fault is decided if $r > h$, where the threshold $h = \max\{r\}|_{t \in H_0}$ is the maximum r obtained during the evaluation of the non-faulty profiles. Hence, the non-faulty profile with data far from the rest of profiles will determine the value for h.

Note from expression (7) in Algorithm 2 that, the farther the new profile data is from the non-faulty data, the lowest the $p(\mathrm{x})$ and the larger the residual r will be.

3 Case Study: Monitoring a Secondary Settler

The present approach is tested using real data from a sensor installed in a secondary settler at Bromma WWTP in Stockholm, Sweden. The sensor measures the suspended solids (SS) concentration as a function of the settler level. The sensor goes from top to bottom of the setter, passing through the clarification and the thickening zone, and measuring the level [m] and the SS concentration [g/L], as shown in Figure 1(a). The profile obtained is called *sludge profile*. A typical sludge profile is shown in Figure 1(b).

Note in Figure 1(a) that we indicate a sludge blanket level, at which there is a jump from lower (less than 0.5 g/L) to higher (above 1 g/L) SS concentration, see Figure 1(b).

The sensor works discontinuously, which means that a new sludge profile is automatically measured after a certain period of time (in minutes). The collected data can be

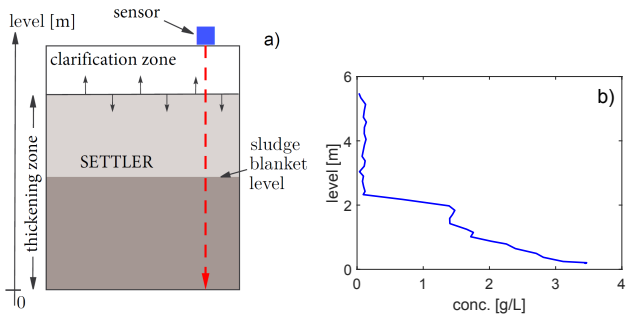


Figure 1. (a) Experiment setup; (b) Typical sludge profile plotted as level vs. SS concentration.

affected by different factors including: changes in the return and/or excess of sludge flow rates, sludge scape, large variations in the influent flow and composition and sensor clogging.

As part of the experiment, two additional measurements were recorded: the level at which the SS concentration is equal to 0.5 g/L (called *fluff level*) and equal to 2.5 g/L (called *sludge level*). We will refer to these levels during the results and discussions of the experiment.

4 Results

Figure 2(a) shows $M = 15$ sludge profiles in non-faulty conditions used for calculating the GMM. Figure 2(b) shows the non-faulty sludge profiles plotted using dots. The highest silhouette value obtained was $S = 0.77$ with $K = 3$, which means that the optimal number of clusters is three, as shown in Figure 2(c). Figure 2(d) shows the contours of the probability density function of the GMM.

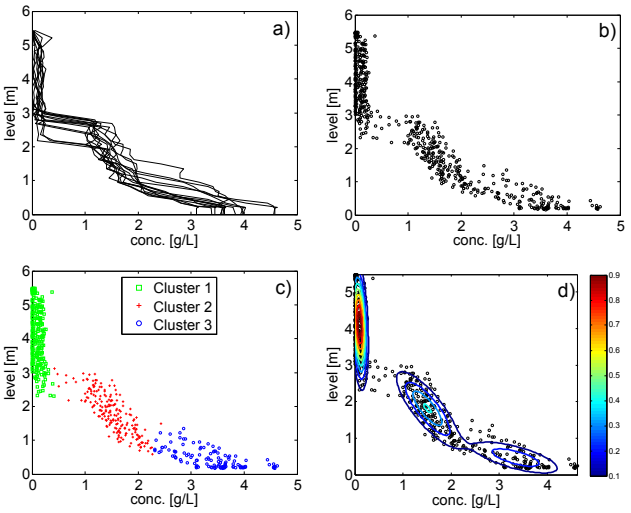


Figure 2. (a) Sludge profiles used to get the GMM; (b) Sludge profile data in (a) plotted using dots; (c) Clusters of the data in (b); (d) Contours of the GMM pdf, color scale indicates the value of the pdf contours.

The GMM parameters π_k, μ_k, and σ_k obtained for the data in Figure 2 are shown in Table 1. There we denote $\mathrm{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$, where $x_1 = \{\text{SS conc.}\}$ and $x_2 = \{\text{level}\}$. Then μ_k and σ_k are

$$\mu_k = \begin{bmatrix} \text{mean}(x_1) \\ \text{mean}(x_2) \end{bmatrix}, \quad \sigma_k = \begin{bmatrix} \text{cov}(x_1,x_1) & \text{cov}(x_1,x_2) \\ \text{cov}(x_2,x_1) & \text{cov}(x_2,x_2) \end{bmatrix},$$

where $k = 1,2,3$ refer to Cluster $1,2,3$, respectively, as shown in Figure 2(c)-(d).

Table 1. GMM parameters

k	Weight (π_k)	Mean (μ_k)	Covariance (σ_k)
1	0.43	$\begin{bmatrix} 0.09 \\ 4.11 \end{bmatrix}$	$\begin{bmatrix} 0.0074 & -0.0223 \\ -0.0223 & 0.7084 \end{bmatrix}$
2	0.34	$\begin{bmatrix} 1.50 \\ 1.82 \end{bmatrix}$	$\begin{bmatrix} 0.1446 & -0.1840 \\ -0.1840 & 0.3550 \end{bmatrix}$
3	0.23	$\begin{bmatrix} 3.34 \\ 0.47 \end{bmatrix}$	$\begin{bmatrix} 0.3612 & -0.1208 \\ -0.1208 & 0.0866 \end{bmatrix}$

The monitoring of the settler was carried out in several trials. As illustration, we present one trial which consisted of 33 days of monitoring, where a new sludge profile was collected every 15 minutes, giving a total of 3168 sludge profiles. In order to see the evolution of the sludge profiles during time, they are shown after 10, 20 and 30 days of running the experiment, as shown in Figure 3(a)-(c), respectively.

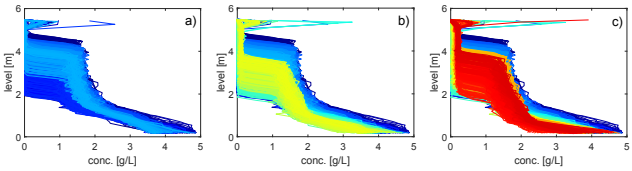


Figure 3. Total of sludge profiles during SST monitoring after: (a) 10 days; (b) 20 days; (c) 30 days.

Figure 4 shows the evolution of the fluff and sludge level, as well as the residual r. The residual r is colored from dark blue (beginning of experiment) to dark red (end of experiment), which correspond to the same range of colors assigned to the sludge profiles shown in Figure 3.

5 Discussions

As mentioned in the case study, a typical sludge profile has an abrupt change in the SS concentration around the sludge blanket level, see the profiles in Figure 2(a). This jump in the SS concentration was captured by the GMM, which classifies the data points before the jump as Cluster 1, and data points after the jump as Cluster 2, as shown in Figure 2(c). Also note that the data points with levels close to zero (bottom of the settler) and with high SS concentration were classified as Cluster 3.

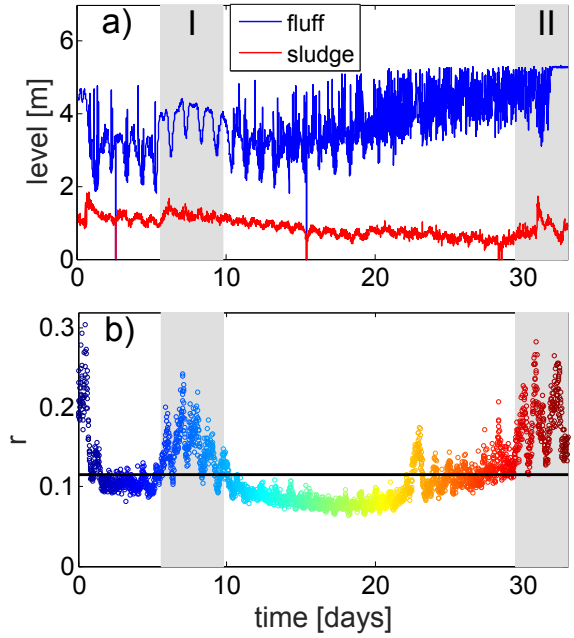


Figure 4. (a) Fluff level (blue line) and sludge level (red line); (b) Residual r (colored dots) and threshold h (black horizontal line). Gray zones refer to Period I and II, see details in Section 5.

From the total of profiles collected during the experiment, we highlight 2 groups of profiles marked as Period I and II in Figure 4. Period I refers to large variations in the influent flow rate, causing fluctuations in the sludge blanket, this effect can also be seen in the oscillatory variation of the fluff level (see Figure 4(a)). The sludge profiles of this period are shown in Figure 5(a). Note in this Figure that several data points at concentrations between 1 and 2 g/L are located far from the pdf contours with high values obtained from non-faulty data, which results in large values for r.

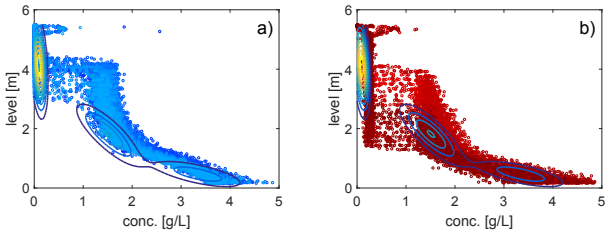


Figure 5. Group of sludge profiles for periods indicated in Figure 4. (a) Period I; (b) Period II. The plots include the contours of the probability density function shown in Figure 2(b).

Another type of events was related to sensor clogging, which began to be detected in profiles during Period II. This clogging event was confirmed by in-situ ocular inspection of the sensor and the existence of floating sludge at the surface level of the settler, promoting sludge escape. Figure 5(b) shows the sludge profiles of this Period. Note in this Figure that several data points are located far from the pdf contours with high values obtained from non-faulty data, particularly at concentrations below 0.5 g/L and between 1 and 2 g/L, which results in large values for r, sometimes even larger than those obtained in Period I.

Note that the data from both periods include outliers. Outliers are defined as sharp changes in the measured values between two successive data. For our case study, outliers in the sludge profiles mean that the measured data is far from the contours obtained with the non-faulty profiles (cf. Figure 2(d)). If there are several outliers in a given sludge profile, it will result in a large value for r. In this study, data correction from outliers was not part of the work. For a process with several events of outliers, the profiles reconstruction could be given by relocating the outliers using the GMM pdf.

Missing data is another possible situation when monitoring profiles. This is, when the amount of data in a given profile is incomplete. In the same way as in the case of outliers, the profiles reconstruction could be given by assigning the missing data using the GMM pdf.

Observe that collecting data from two sensors measuring the same process, the total set of data from each sensor will be different, resulting that each sensor will have a unique probability density function. This means that the present methodology has an important advantage, since is not just applied to specific sensors or processes but to sensors or processes from diverse areas.

A possible application for the current approach is to use the residual value r as a tool for a control action. In this way, it would be possible to formulate different control strategies based on, for example, changes in the recycle flow rate of the WWTP, in order to keep the new sludge profiles as similar as possible to the non-faulty profiles.

6 Conclusions

A GMM-based approach for monitoring and fault detection of the sludge profiles in a SST working in a WWTP has been proposed. Using a set of non-faulty profiles, the aim was to obtain a non-faulty region (SS concentrations, SST height) defined by a pdf via the GMM method. This pdf is then used to evaluate new profiles and detect possible abnormal profiles. Results obtained with real data implementation suggest that this method could help to monitor the performance of a SST.

Acknowledgment

The authors acknowledge funding support under the European Union's Seventh Framework Programme managed by the Research Executive Agency (REA), Grant Agreement N.315145 (Diamond). Funding from Käppala Association, Syvab and Stockholm Water Company, Foundation for IVL Swedish Environmental Research Institute and the Swedish Water and Wastewater Association is gratefully acknowledged.

References

Christopher Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007. ISBN 0387310738.

Ratko Grbić, Dino Kurtagić, and Dražen Slišković. Stream water temperature prediction based on Gaussian process regression. *Expert Systems with Applications*, 40(18):7407–7414, 2013a. doi:10.1016/j.eswa.2013.06.077.

Ratko Grbić, Dražen Slišković, and Petr Kadlec. Adaptive soft sensor for online prediction and process monitoring based on a mixture of Gaussian process models. *Computers & Chemical Engineering*, 58:84–97, 2013b. doi:10.1016/j.compchemeng.2013.06.014.

Koen Grijspeerdt and Willy Verstraete. Image analysis to estimate the settleability and concentration of activated sludge. *Water Research*, 31(5):1126–1134, 1997. doi:10.1016/s0043-1354(96)00350-8.

Juš Kocijan and N. Hvala. Sequencing batch-reactor control using Gaussian-process models. *Bioresource Technology*, 137:340–348, 2013. doi:10.1016/j.biortech.2013. 03.138.

Ben Li and M.K. Stenstrom. Research advances and challenges in one-dimensional modeling of secondary settling tanks – a critical review. *Water Research*, 65:40–63, 2014. doi:10.1016/j.watres.2014.07.007.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)*. The MIT Press, 2012. ISBN 0262018020.

G. Olsson, B. Carlsson, J. Comas, J. Copp, K. V. Gernaey, P. Ingildsen, U. Jeppsson, C. Kim, L. Rieger, I. Rodríguez-Roda, J.-P. Steyer, I. Takács, P. A. Vanrolleghem, A. Vargas, Z. Yuan, and L. Åmand. Instrumentation, control and automation in wastewater – from London 1973 to Narbonne 2013. *Water Science & Technology*, 69(7):1373, 2014. doi:10.2166/wst.2014.057.

Michael A. Osborne, Roman Garnett, Kevin Swersky, and Nando De Freitas. Prediction and fault detection of environmental signals with uncharacterised faults. In *Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*, 2012.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning series)*. The MIT Press, 2005. ISBN 026218253X.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi:10.1016/0377-0427(87)90125-7.

Adama Traoré, Stéphane Grieu, Frédérik Thiery, Monique Polit, and Jésus Colprim. Control of sludge height in a secondary settler using fuzzy algorithms. *Computers & Chemical Engineering*, 30(8):1235–1242, 2006. doi:10.1016/j.compchemeng.2006.02.020.

Chang Kyoo Yoo, Sang Wook Choi, and In-Beum Lee. Adaptive modeling and classification of the secondary settling tank. *Korean Journal of Chemical Engineering*, 19(3):377–382, 2002. doi:10.1007/bf02697143.

Jie Yu. A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes. *Chemical Engineering Science*, 68(1): 506–519, 2012. doi:10.1016/j.ces.2011.10.011.

Jesús Zambrano, Oscar Samuelsson, Tatiana Chistiakova, Hongbin Liu, and Bengt Carlsson. Gaussian process regression for monitoring a secondary settler. In *2nd New Development in IT and Water*, Rotterdam, The Netherlands, 2015.

Hongyan Zhu, Shuo Chen, and Chongzhao Han. Fusion of gaussian mixture models for possible mismatches of sensor model. *Information Fusion*, 20:203–212, 2014. doi:10.1016/j.inffus.2014.02.002.

Industrial Model Validation of a WWT Bubbling Fluidized Bed Incinerator

Souad Rabah[1,2,3,4] Rodrigo O. Brochado[2,4] Hervé Coppier[1,2] Mohammed Chadli[1,2] Nesrine Zoghlami[4] Mohamed Saber Naceur[4] Sam Azimi[5] Vincent Rocher[5]

[1]MIS Laboratory, 80000 Amiens, France.
[2]ESIEE-Amiens, 80082 Amiens, France. {COPPIER,RABAH_S}@esiee-amiens.fr
[3]University of Picardie Jules Verne,(UPJV) 80039 Amiens, France. mohammed.chadli@u-picardie.fr
[4]LTSIRS Laboratory, ENIT,Tunis El Manar University,1002 Tunis,Tunisia
[5]Direction of Developement and Prospective, SIAAP, 92700 Colombes, France.

Abstract

The environmental concern has significantly raised and specially in the case of bioprocess industries where new standards are more and more strict. In this context, areas of new research have been developed to enhance biological treatment processes productivity and reduce emission of toxic substance. To ensure the control of the incineration variables, a describing model should be determined. In our previous work, we presented multi variable identification results relative- to SIAAP incineration process. This paper concerns model validation of a fluidized-bed incineration furnace by different quality criteria. In this case, our focus is on production phase defined by two incineration modes. Thus, the observed data of each model is compared with the predicted data using quality criteria.

Keywords: validation methods, fluidized-bed furnace, subspace state-space system N4SID

1 Introduction

Knowing well a dynamic system is a good way to develop its robust efficient control. It is useful for precise weather forecast, general data predictions, model-based simulation and others. For this reason, modelling process and its validation need great attention. Over the past years, new identification methods have been developed to achieve better results in some specific systems, as it can be seen in (Grossmann et al., 2009). Other traditional methods are also developed and adopted, as the N4SID algorithm (Van Overschee and De Moor, 2012; Rabah et al., 2016), that still has satisfactory results in system identifications (Grossmann et al., 2009; Panday et al., 2009; Kojio et al., 2014). An identification process results in a dynamic model which is believed to be the best approximation of certain real process.

However, even using tested and approved methods, a system identification may not give what is expected from its model. It can diverge in peak values, up or down trends, model time constant and more.In consequence, one or more validation methods must be applied on the identified model to analyses its fit with the real process.

In this paper, model validation of a fluidized-bed incineration furnace by six quality criteria is performed. The furnace is owned by SIAAP (Rabah et al., 2016; Mailler et al., 2014) that provides raw measured data to be processed. The measured temperature of each part of the furnace as well as other signals useful for modelling are provided in a thirty-second sample time. Prediction data of resultant model are then compared to observed data for each incinerator sub model. The validation methods adopted are known as NRMSE, LCE and NMSE, AME, MSE and MSDE defined in (Hauduc, 2011). Finally, the results are analyzed and compared to each other. Validation methods are also used for on-line identification methods.

This paper is organized as follows: in the next section SIAAP is presented. The identification method and its results are presented in Section 3. The validation methods adopted for this paper are defined in Section 4. The validation results for each model are presented in section 5.

2 Industrial Process Description

2.1 SIAAP Waste water and sewage Sludge Treatment Process

The SIAAP is a French public institution. It was founded in 1971 by the Council of Paris in order to perform the wastewater treatment (WWT) of Greater Paris due to the poor quality of the city's rivers at that time. Nowadays, SIAAP performs the transport, storage, management and purification of wastewater of 180 communes reaching almost nine million users. It can purify more than 2,5 million m^3 (Mailler et al., 2014) of wastewater each day with its 6 treatment stations.

In Seine Centre Plant, where the studied sludge incinerator furnace is located, WWT flux can get up to 2,8 m^3 per second in dry weather. That infrastructure allows to change treatment process and can purify up to 12 m^3 of wastewater per second. In this site, there are two different ways of treatment performed: wastewater treatment and sludge treatment.

The WWT in Seine Center Plant is divided in three

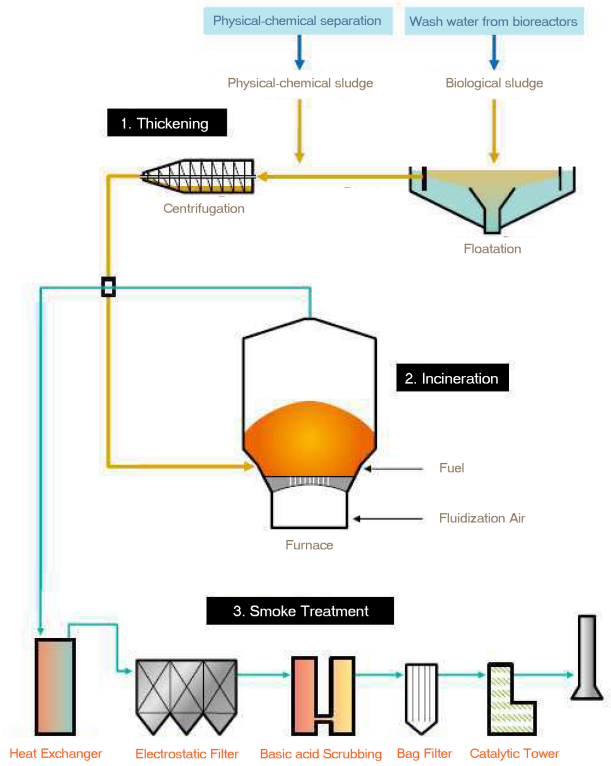


Figure 1. Sludge Treatment Diagrams.

stages. The preliminary treatment allows eliminate big particles like sand or fats. The second one, decantation, allows the elimination of much part of colloids, suspended solids and orthophosphates. The third and last one, biological treatment, allows the elimination of carbon and nitrogen pollution.

In Figure 1 we present sludge treatment part. Sludge is the result of WWT. First of all, sludge is treated passing by floatation and dehydration process in order to control the water concentration in the sludge, in a way that it does not disturb its incineration. In the first one air and polymer are injected to promote in the process of water and sludge separation and in the second one another polymer is injected to sludge thickening. Then, sludge is injected inside the fluidized-bed furnace to incineration. The smoke produced is then treated by five different processes, aimed to control smoke temperature and pollution level. Hot air issued of the smoke temperature treatment represent the fluidised air in order to maintain furnace temperature.

Relative data from the sludge treatment process of Seine Centre Plant is given by SIAAP for the development of this research.

2.2 Incineration Process

The incineration in the Waste Water Treatment Plants (WWTPs) is a thermal recycling relative to sewage sludge treatment. The purpose of this technologies is to solve hygiene problems while the WWTPs effluents are considered as a source of contamination (Mailler et al., 2014). Despite the complexity of the combustion mechanism involved and the strongly non-linearity, the use of this technology has increased thanks to new thermal recovery strategies (Martins et al., 2014; Tong et al., 2012; Ravelli et al., 2008; Khiari et al., 2008; Hadavand et al., 2008; Li and Li, 2016).

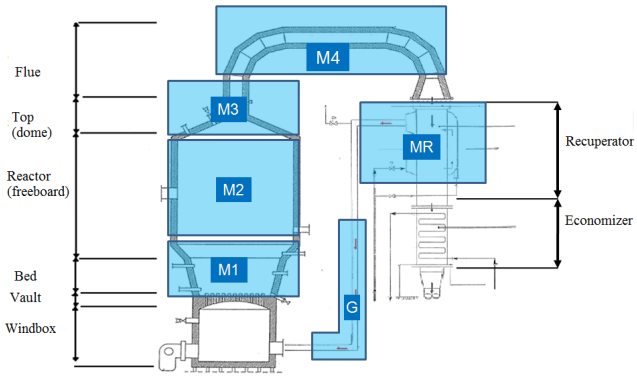


Figure 2. The incineration process sketch.

Figure 2 presents the considered incineration process which is composed by two principal process: the furnace (incinerator) and the heat Exchanger.

2.2.1 The Furnace

The Seine center furnace is called Pyrofluid incinerator. it is based on the bubbling fluidized bed technology. This system is consist of 6 important parts as shown in Figure 2:

- **Windbox**: it ensures the preheated fluidised air transfer to the Bed, and has a major role in the boot process.
- **Vault**: it ensures uniform distribution of air thanks to tuyeres.
- **Fluidised bed**: it is composed of sand as inert particles. The Injection fluidised air in the bed ensure uniform turbulence.
- **Freeboard**: its volume allows the complete combustion of organic material. In the reactor, the temperature should be maintained at **T2S** = 850 °C for at least 2 seconds.
- **Dome**: at this level, NH_4OH is injected in order to reduce NO_x toxicity and water to avoid high temperature.
- **Flue**: it ensures fume transport to the heat exchanger.

The load temperature of this process is from 800 °C to 950 °C and it ensures the total combustion of sludge in just few seconds. This technique reduces the volumes of the

incinerated waste by 90% and the mass by 70% (Rabah et al., 2015).

2.2.2 The Heat Exchanger

In fact, the heat exchanger is composed by two exchangers blocks: recuperator and economizer.

In order to ensure an optimal energy recovery, the thermal energy resulting of combustion phenomenon is recycled in the process thanks to the recuperator that preheats the fluidizing air and the economizer, allows the water heater for other needs at the station.

3 Identification Process Overview

The fluidized bed incineration is one of the key innovative technologies in the field of sewage sludge incineration treatment but also the most complex one because of the difficulty of establishing a representative model of the process in order to improve the system efficiency. However, to understand the dynamics of fluidized bed is necessary but not sufficient to improve the performance of this technology. The development of numeric identification methods provide tools to model the dynamics of these complex systems. These models are usually used in order to improve control and guarantee optimal performance.

The furnace is driven by external excitation, these are the system "Inputs". The reaction of the process which is measured by the sensors is called system "Output". The objective in this step is to synthesize a mathematical model that describes the system reaction, this model will be able to predict the output of the system.

The incineration process is decomposed into 6 sub-models interconnected between them as presented in Figure 3.

3.1 Sludge Incineration sub-models: Input-output interaction.

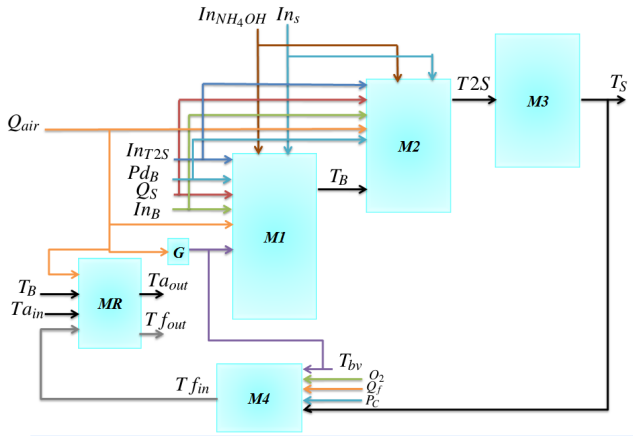


Figure 3. Sludge Incineration: input-output Sub-model.

- $M1$ - Bed Model (T_B).
- $M2$ - Reactor Model ($T2S$).
- $M3$ - Top Reactor Model (T_S).
- $M4$ - Flue model(Tf_{in}).
- MR - Recuperator Model (Tf_{out},Ta_{in})
- G - Model of thermal dissipation (T_{bv})

3.2 Identification Strategy

The identification is to find mathematical models of systems based on experimental data (black box model) and available knowledge (Grey box model or semi-physical model). These models must provide an approximation of the considered system in order to calculate the physical parameters or design simulation algorithms, monitoring or control. The classic approach is to formalize the available knowledge, to collect experimental data and estimate the structure, the parameters, and end up with validation. The identification methods are generally based on minimizing of the "prediction error". In other words, these techniques establish a model by reducing prediction error of the future output Y_i while considering past output Y_0 and input U_0 and U_i. In this case, we present validation result of identification model based on the subspace state space system identification method (Favoreel et al., 2000; Hachicha et al., 2014; Jamaludin et al., 2013; Van Overschee and De Moor, 1994; Viberg, 1995). All models are presented as a linear discrete time invariant state space model (LTI) (Pekpe, 2004; Van Overschee and De Moor, 2012):

$$\begin{cases} x_{k+1} = Ax_k + Bu_k + v_k \\ y_k = Cx_k + Du_k + \omega_k \end{cases} \quad (1)$$

Where $x_k \in \mathbb{R}^n$ is the state vector at discrete time instant k, $u_k \in \mathbb{R}^m$ is the input, $y_k \in \mathbb{R}^l$ is the output while $v_k \in \mathbb{R}^n$ and $\omega_k \in \mathbb{R}^l$ are additional unobserved signals sequences, v_k is the measurement noise and ω_k the process noise. The identification result was presented in (Rabah et al., 2016).

4 Validation Methods

The validation of identified models uses the relative reference-model criteria that compares the residuals of the identified model to the ones of a simplistic and reference model. In order to know error greatness, the absolute criteria will also be used. According to (Hauduc, 2011), neither the relative error criteria nor the graphical methods are considered. In fact, the relative error criteria does not offer a relation between two models as the relative reference-model and the graphical methods don't state a numerical result of its method, which resulting in a qualitative method.

4.1 Absolute Criteria

In this paper, absolute criteria approach avoid error compensation. And there is an optimal cost value of zero in this approach. Error dimensions are given in the square of the measured temperature unit.

The AME method, defined by (2), denotes the absolute maximum error. It is used to know the model's behaviours, but it does not describe the real process very well. The MSE method, defined by (3), emphasizes high errors but it does not have a reference, which is difficult to decide if its result is satisfactory or not. The MSDE method, defined by (4), is calculated based on two time steps. It denotes peak's timing errors.

$$AME(^{\circ}\mathrm{C}) = max(|O_i - P_i|) \tag{2}$$

$$MSE(^{\circ}\mathrm{C}^2) = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2 \tag{3}$$

$$MSDE(^{\circ}\mathrm{C}^2) = \frac{1}{n-1}\sum_{i=1}^{n}((O_i - O_{i-1}) - (P_i - P_{i-1}))^2 \tag{4}$$

4.2 Relative reference-model criteria

The NRMSE method, defined by (5), emphasizes larger errors in the identified model and can be compared to other methods to indicate the influence of this type of error. One could state that larger errors would not be emphasized by this method, but as seen in (5), the square root is present in both numerator and denominator. The denominator root underlines errors when the observed data is much closer to its mean than to predicted data.

The LCE method, defined by (6), emphasizes very low magnitude errors. Natural logarithm in both numerator and denominator puts all residuals in a lower scale in a way that smaller residuals have influences in results closer to the influence of greater ones. Then, greater errors continue to be considered in LCE criteria, but with lower emphasis.

Finally, NMSE method, defined by (7), also emphasizes high errors, but without a square root. It seems that greater residuals will give lower criteria values, but results of this criteria tends to be more elevated than NRMSE - i.e. better results - when the observed data is closer to its mean than to predicted data.

All these methods avoid errors compensations and have an optimal cost value of 1. Values close to zero indicate that the identified model is not better than the reference model, which is all in this case a constant equal to observed data mean. Negative values indicate a model worse than the reference model. Lower possible value is negative infinity.

$$NRMSE(\%) = 1 - \frac{\sqrt{\sum_{i=1}^{n}(O_i - P_i)^2}}{\sqrt{\sum_{i=1}^{n}\left(O_i - \bar{O}\right)^2}} \tag{5}$$

$$LCE(\%) = 1 - \frac{\sum_{i=1}^{n}(\ln O_i - \ln P_i)^2}{\sum_{i=1}^{n}\left(\ln O_i - \overline{\ln O}\right)^2} \tag{6}$$

$$NMSE(\%) = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}\left(O_i - \bar{O}\right)^2} \tag{7}$$

5 Results and Discussions

This section presents the validation of each sub model identified using a subspace method. The recuperator part that comes after the flue, is divided in two models. One analyses the fluidised air and the other deals with the fume temperature. The models are called MR_{AirOut} and $MR_{SmokeOut}$, respectively. The output signal of each model is the temperature of its respectively furnace section. validation algorithms are developed using MATLAB interface in order to get NRMSE, NMSE, MSE, LCE, AME and MSDE results.

Tables 1 and 2 show results of each validation method for each sub model for absolute criteria and relative reference-model criteria, respectively. Two different furnace operation modes are presented: one with reactor fuel injection and the other without fuel injection.

Considering Section 4, absolute criteria has subjective results, having different quality standard values depending on very specific system characteristics. However, these methods could be useful for superficial analysis. It is seen in Table 1 that the models present tolerable AME results, except for the model $MR_{SmokeOut}$, knowing that temperature in the furnace can reach values in the order of $900^{\circ}C$. MSE results cannot be analysed singly due to its subjectivity, but the results of the same model in both operation modes can be compared. It is seen that M1, M2, M3 and $MR_{SmokeOut}$ have better MSE values in operation without reactor fuel injection than operation with fuel.

Table 1. Absolute Criteria Results of Each Validation Method with and without Reactor Fuel Injection.

	Sub Model	*AME*	*MSE*	*MSDE*
Without Fuel	*M1*	4.40	2.70	0.07
	M2	10.12	7.31	1.51
	M3	6.00	2.00	0.16
	M4	28.85	15.32	0.08
	G	6.43	8.74	0.05
	MR_{AirOut}	14.31	38.63	0.02
	$MR_{SmokeOut}$	12.59	25.65	0.05
With Fuel	*M*1 / M2	22.25	18.44	2.65
	*M*3	33.19	53.08	0.52
	*M*4	10.29	2.98	0.08
	G	8.49	1.83	0.05
	MR_{AirOut}	21.50	16.12	0.05
	$MR_{SmokeOut}$	52.47	142.88	0.34

5.1 Relative Reference-Model Criteria Results

At first, we note that the values have a sorting relation, in a way that if all the NRMSE values are sorted in ascend-

Table 2. Relative Reference-Model Results of Each Validation Method with and without Reactor Fuel Injection.

	Sub Model	*NRMSE*	*LCE*	*NMSE*
Without Fuel	*M1*	55.24	79.95	79.96
	M2	53.20	78.07	78.09
	M3	83.65	97.31	97.34
	M4	73.15	92.59	92.79
	G	82.53	96.96	96.95
	MR_{AirOut}	73.30	92.77	92.87
	$MR_{SmokeOut}$	66.68	88.76	88.90
With Fuel	*M1 / M2*	57.97	82.34	82.33
	M3	48.29	73.62	73.26
	M4	61.69	85.44	85.32
	G	85.73	98.00	97.96
	MR_{AirOut}	73.73	93.28	93.10
	$MR_{SmokeOut}$	68.35	90.68	89.98

ing or descending order, the other two criteria results will follow the same order. Thus, the bigger NRMSE values are, the bigger LCE or NMSE values will be. It does not imply a linear relation among these criteria. It just makes it clear the different emphasis that each method gives to residuals types.

Considering this point, it can also be said that all the models have similarities in the type of the error. For example, M1 with fuel has a NRMSE result of 42,01% and LCE result of 65,92% and M1 without fuel has a NRMSE result of 55,24% and LCE result of 79,95%.

Moreover, it can also be suggested that some data are correlated, as it has the same type of error. It means that the inputs and outputs signal of one sub model may be related to the inputs and outputs signal of another one.

Each system has different response curves, like swinging and noisy curves, slowly and soft curves, periodical curves and more. Like each validation method lay emphasis on a kind of residual, each system should adopt a different validation method. For this paper, fluidized-bed furnace system is identified and should be validated. As the temperature curves have so many peaks, peak timing validation methods like MSDE (Hauduc, 2011) should be chosen. Looking at the residuals values, it is noted that in most cases that their values are not so high. Thus, a validation method for low magnitude residuals may be adopted. It is shown in Table 2 that the LCE results for most of models are excellent.

As for on-line identification, validation methods need to be extremely secure in some cases. All the precautions must be taken to avoid inefficient or unstable models. To prevent this situation, redundancy methods should be performed. In a general view, a method that takes in account high and low magnitudes residuals could do well. As a suggestion, LCE and NRMSE methods should be adopted simultaneously. This on-line identified model validation has not yet been proven, but it is a subject of next studies.

6 Conclusions

This paper investigates model validation of a fluidized-bed incineration furnace by different quality criteria. The model of each part of the furnace is identified with the N4SID method using MATLAB function. The three methods adopted present good results. Regarding the type of error of the dynamic system being studied, low magnitude error methods should be performed. LCE method presents excellent results. For on-line model identification, both LCE and NRMSE are suggested.

Notation

SIAAP	Paris urban area waste-water treatment authority.
NRMSE	Normalized Root Mean Square Error.
LCE	Logarithmic residuals Comparison errors.
NMSE	Normalized Mean Square Error.
AME	Absolute Maximum Error.
MSE	Mean Square Error.
MSDE	Mean Square Derivative Error.
O_i	Observed model data.
P_i	Predicted model data.
$\bar{O}$	Observed data mean
n	The number of validation data.
$T2S$	Reactor Temperature (Temperature of 2s) (°C).
T_S	Temperature in the top of the incinerator (°C).
T_B	Bed Temperature (°C).
T_{bv}	Windbox Temperature (°C).
In_{T2S}	Reactor fuel injection (l/s).
In_B	Bed fuel injection (l/s).
Q_S	Sludge flow (l/s).
In_{NH_4OH}	NH_4OH injection (l/s).
In_s	Water top reactor injection (l/s).
Q_{air}	Fluidizing air flow (Nm^3)
Pd_B	Bed differential pressure($mbar$)
Tf_{in}	Recuperator input fume temperature (°C).
Tf_{out}	Recuperator output fume temperature (°C).
Ta_{in}	Recuperator input Air temperature (°C).
Ta_{out}	Recuperator output Air temperature (°C).
O_2	Oxygen measure (%).
Q_f	Flue gas flow (Nm^3).
P_C	Flue pressure (Nm^3).

Acknowledgment

This study is being conducted as part of Axe 3 of the program MOCOPEE[1] (phase 1. 2014 - 2017). The authors would like to thank MOCOPEE program for the financial support.

References

W. Favoreel, B. De Moor, and P. Van Overschee. Subspace state space system identification for industrial processes. *Journal of process control*, 10(2):149–155, 2000.

C. Grossmann, C. N. Jones, and M. Morari. System identification via nuclear norm regularization for simulated moving bed processes from incomplete data sets. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pages 4692–4697. IEEE, 2009.

[1] MOCOPEE: MOdélisation, Contrôle et Optimisation des Procédés d'Epuration des Eaux: www.mocopee.com

S. Hachicha, M. Kharrat, and A. Chaari. N4SID and MOESP algorithms to highlight the ill-conditioning into subspace identification. *International Journal of Automation and Computing*, 11(1):30–38, 2014. ISSN 1476-8186.

A. Hadavand, A. A. Jalali, and P. Famouri. An innovative bed temperature-oriented modeling and robust control of a circulating fluidized bed combustor. *Chemical Engineering Journal*, 140(1):497–508, 2008. doi:10.1016/j.cej.2007.11.032.

H. Hauduc. *Modeles biocinétiques de boues activées de type ASM: Analyse théorique et fonctionnelle, vers un jeu de parametres par défaut*. PhD thesis, Laval University, 2011.

I. W. Jamaludin, N. A. Wahab, N. S. Khalid, S. Sahlan, Z. Ibrahim, and M.F. Rahmat. N4sid and moesp subspace identification methods. pages 140–145. IEEE, 2013. doi:10.1109/CSPA.2013.6530030.

B. Khiari, F. Marias, F. Zagrouba, and J. Vaxelaire. Transient mathematical modelling of a fluidized bed incinerator for sewage sludge. *Journal of Cleaner Production*, 16(2):178–191, 2008. doi:10.1016/j.jclepro.2006.08.020.

J. Kojio, H. Ishibashi, R. Inoue, S. Ushida, and H. Oku. Mimo closed-loop subspace model identification and hovering control of a 6-dof coaxial miniature helicopter. In *SICE Annual Conference (SICE), 2014 Proceedings of the IEEE*, pages 1679–1684. IEEE, 2014.

S. Li and Y. Li. Model predictive control of an intensified continuous reactor using a neural network wiener model. *Neurocomputing*, 2016. doi:10.1016/j.neucom.2015.12.048.

R. Mailler, J. Gasperi, V. Rocher, S. Gilbert-Pawlik, D. Geara-Matta, R. Moilleron, and G. Chebbo. Biofiltration vs conventional activated sludge plants: what about priority and emerging pollutants removal? *Environmental Science and Pollution Research*, 21(8):5379–5390, 2014.

M. A. F. Martins, A. C. Zanin, and D. Odloak. Robust model predictive control of an industrial partial combustion fluidized-bed catalytic cracking converter. *Chemical Engineering Research and Design*, 92(5):917–930, 2014. doi:10.1016/j.cherd.2013.08.005.

R. Panday, B. D. Woerner, J. C. Ludlow, L. J. Shadle, and E. J. Boyle. Linear system identification of a cold flow circulating fluidized bed. *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering*, 223(1):45–60, 2009.

K. M. Pekpe. *Identification par les techniques des sous-espaces-application au diagnostic*. PhD thesis, Institut National Polytechnique de Lorraine-INPL, 2004.

S. Rabah, H. Coppier, M. Chadli, N. Zoghlami, and M. S. Naceur. Régulation multi-variable pour un incinérateur à lit fluidisé circulant: approche lmi. In *6èmes Journées Doctorales / Journées Nationales MACS*, 2015.

S. Rabah, H. Coppier, M. Chadli, S. Azimi, V. Rocher, D. Escalon, N. Zoghlami, and M. S. Naceur. Multi-variable industrial processes identification: Case of bubbling fluidized bed sewage sludge incinerator. In *Control and Automation (MED), 2016 24th Mediterranean Conference on*, pages 803–808. IEEE, 2016.

S. Ravelli, A. Perdichizzi, and G. Barigozzi. Description, applications and numerical modelling of bubbling fluidized bed combustion in waste-to-energy plants. *Progress in Energy and Combustion Science*, 34(2):224–253, 2008. doi:10.1016/j.pecs.2007.07.002.

H. Tong, X. Zhao, and G. Liang. Design and simulation of fuzzy decoupling for combustion process of circulating fluidized bed boiler. In *Advances in Electronic Engineering, Communication and Management Vol. 2*, pages 137–143. Springer, 2012.

P. Van Overschee and B. De Moor. N4sid: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30(1):75–93, 1994.

P. Van Overschee and B. L. De Moor. *Subspace identification for linear systems: Theory - Implementation - Applications*. Springer Science & Business Media, 2012.

M. Viberg. Subspace-based methods for the identification of linear time-invariant systems. *Automatica*, 31(12):1835–1851, 1995.

Simulation of Oil Production in a Fractured Carbonate Reservoir

Nora Cecilie Ivarsdatter Furuvik Britt M. E. Moldestad

Department of Process, Energy and Environmental Technology, University College of Southeast Norway, Norway,
{nora.c.i.furuvik, britt.moldestad}@usn.no

Abstract

CO_2-EOR is an attractive method because of its potential to increase the oil production from matured oilfields, at the same time reduce the carbon footprint from the industrial sources. The field response to the CO_2-EOR technique depends on the petrophysical properties of the reservoir. Carbonate reservoirs are characterized by low permeability and strong heterogeneity, causing significant amounts of water and CO_2 to be recycled when CO_2 is re-injected into the reservoir. Naturally fractured carbonate reservoirs have low oil production, high water production, early water breakthrough and high water cut. This study focuses on the oil production and the CO_2 recycle ratio in naturally fractured carbonate reservoirs, including near-well simulations using the reservoir software Rocx in combination with OLGA. The simulations indicate that closing the fractured zone causes delayed water breakthrough and dramatically reduced water cut, resulting in improved oil recovery as well as lower production and separation costs.

Keywords: CO_2-EOR, fractured carbonate reservoirs, inflow control, near well simulation

1 Introduction

Deep geologic injection of supercritical carbon dioxide (CO_2) for enhanced oil recovery (EOR) plays an important role in the sequestration of CO_2 to minimize the impact of CO_2-emissions due to global warming (Ettehadtavakkol *et al*, 2014; Hill *et al*, 2013). CO_2-EOR refers to the oil recovery technique where supercritical CO_2 is injected into the oil reservoir to stimulate the oil production from depleted oilfields. The CO_2 mixes with the stranded oil and change the oil property, making the immobile oil mobile and producible (Ettehadtavakkol *et al*, 2014).

The efficiency of the CO_2-EOR technique greatly depends on the petrophysical properties of the reservoir (Ettehadtavakkol *et al*, 2014; Tarek, 2014). In carbonate reservoirs, the petrophysical properties generally are controlled by the presence and the distribution of naturally fractures. Fractures are high permeability pathways for fluid migration in a low permeability rock matrix (Fitch, 2010; Moore, 1989). Oil recovery from carbonate reservoirs with fractures are challenging compared to oil recovery from other reservoirs, as the fluids preferably will flow through the high permeable fractures. The result is poor sweep efficiency and potentially low oil recovery, due to very early water breakthrough (Haugen, 2010).

Most carbonate reservoirs are naturally fractured, causing significant amounts of water and CO_2 to be produced together with the main stream from the production well during the CO_2-EOR process. (Fitch, 2010; Ettehadtavakkol *et al*, 2014). For the oil companies this is both economic, operational and environmental challenging. High demands and rising oil prices has increased the focus on new inflow technology to improve oil recovery from low recovery oilfields (Tarek, 2014). The breakthrough of water and CO_2 can be limited by installing Autonomous Inflow Control Valves (AICV) in the inflow zones in the well. The AICV will automatically shut off the production of water and CO_2 from one specific zone in the well, but at the same time continue the production of oil from other zones. The AICV can replace the conventional Inflow Control Devices (ICD) installed in a well (Brettvik, 2013).

This study focuses on CO_2-EOR in naturally fractured carbonate reservoirs, including simulations of oil production from an oil-wet reservoir. Both ICD and AICV completion were simulated in order to study the benefits of the AICV technology. The simulations are carried out using commercial reservoir simulation software.

2 CO_2-EOR

CO_2-EOR is a technique that involves injection of supercritical CO_2 into underground geological formations, or deep saline aquifers. The goal is to revitalize matured oilfields, allowing them to produce additional oil. CO_2 is highly soluble in oil and to a lesser extent in water. As CO_2 migrates through the reservoir rock, it mixes with the residual oil trapped in the reservoir pores, enabling the oil to slip through the pores and sweep up in the flow from the CO_2-injection well towards the recovery well. (Hill *et al*, 2013) The principle of CO_2-EOR is shown in Figure 1.

When supercritical CO_2 and oil mix, a complicated series of interactions occur wherein the mobility of the crude oil is increased. These interactions involve

reduction in the interfacial tensions and the capillary pressure between the oil and the water phase. Injection of CO_2 into the oil formation changes the oil physical properties in two ways, leading to enhanced oil recovery. First, the oil viscosity is reduced so that the oil flows more freely within the reservoir. Then, a process of dissolution occur thereby causing swelling of the oil, resulting in expansion in oil volume which means that some fluid have to migrate. The amount of swelling depends on the reservoir pressure and temperature, the hydrocarbon composition and the physical properties of the oil (Hill *et al*, 2013; Pasala, 2010, NRG Energy, 2014; Ghoodjani *et al*, 2011).

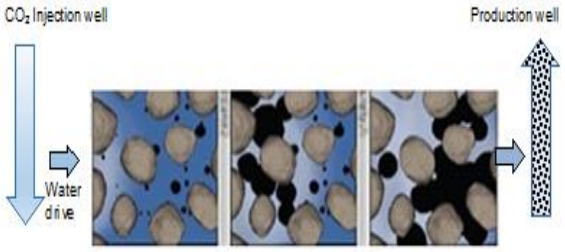


Figure 1. Principle of CO_2-EOR (Oil and Gas 360, 2016).

3 Carbonate reservoirs

More than 60 % of the world's oil resources occur in carbonate rocks (Fitch, 2010). Although carbonate reservoirs contain a majority of the oil reserves, only small amounts of the oil production worldwide come from these reservoirs (Fitch, 2010). Generally, carbonate reservoirs are characterized by complicated pore structures and strong heterogeneity. The heterogeneity of carbonate reservoirs is the result of a complex mineral composition and a complex rock texture. The heterogeneity is one of the main reasons causing low oil recovery from carbonates, as it contributes to highly variability in the petrophysical properties within small sections of the reservoir (Fitch, 2010; Moore, 1989).

3.1 Petrophysical properties of carbonate reservoirs

The petrophysical properties are controlled by the presence and the distribution of open fractures. Most carbonate reservoirs have a dual character of rock matrix and natural fractures. Fractures are discontinuities in the rock appearing as breaks in the natural sequence. The orientation of the fracture can be anywhere from horizontal to vertical, as illustrated in Figure 2. The fractured corridors exist in all scales, ranging from microscopic cracks to fractures of ten to hundreds of meters in width and height. This results in greatly variable permeability in carbonate reservoirs, from values less than 0.1 mD in cemented carbonates to over 10 000 mD in fractures and have a considerable impact on oil production (Fitch, 2010; Moore, 1989).

Porosity is another important parameter affecting the oil recovery as it is a result of the secondary processes involving compaction and cementation of the sediments, and is controlled by the original grain shape and grain size distribution. Porosity in carbonate reservoirs varies from 1 % - 37 % (Fitch, 2010).

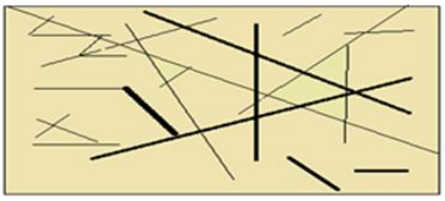


Figure 2. Fractures in reservoir.

Wettability of the reservoir describes the preference for the rock matrix to be in contact with one certain fluid phase over another. The reservoirs can be either water-wet or oil-wet (Satter *et al*, 2007). An oil-wet reservoir has higher affinity for the oil phase than for the water phase, oil will occupy the smaller pores and preferably stick to the grain surface in the larger pores. In oil-wet reservoirs, attractive forces between the rock and the fluid draw the oil into the smaller pores. While repulsive forces cause the water to remain in the center of the largest pores. The opposite condition is water-wet reservoir, in which the pore surface prefers contact with the water phase and water absorbs into the smaller pores. The wetting phase fluid often has low mobility, while the non-wetting fluid is more mobile and especially at large non-wetting phase saturations (Schlumberger, 2007; Ahmed, 2013, International Human Resources Development Corporation, 2016). A great majority of carbonate reservoirs tend to be oil-wet. Extensive research work on wettability for carbonate reservoir rocks confirms that carbonates exhibit significantly more oil-wet character than water-wet character. Performed contact angle measurements show that 15 % of carbonates are strongly oil-wet (θ=160°-180°), 65 % are oil-wet (θ=100°-160°), 12 % are intermediate-wet and 8% are water-wet (Esfahani *et al*, 2004). Evaluations of wettability for the carbonate rock samples, using relative permeability curves and Amott tests conclude that the carbonate reservoirs investigated ranges from intermediate-wet to oil-wet (Esfahani et al, 2004). Figure 3 illustrates the difference between a water-wet and an oil-wet reservoir rock.

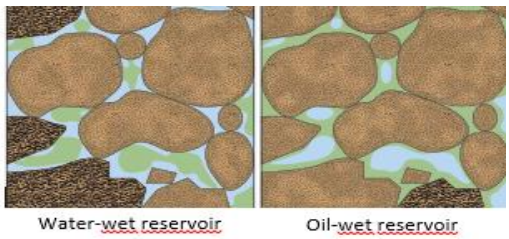


Figure 3. Wetting in pores (Schlumberger, 2007).

Presumed petrophysical properties of carbonate reservoirs are presented in Table 1.

Table 1. Petro physical properties of carbonate reservoirs (Fitch, 2010; Moore, 1989).

Porosity	*Permeability*	*Permeability in fracture*	*Wettability*
0.01-0.3	0.7-130 mD	Large	Intermedia te-wet to strongly oil-wet

3.2 CO_2-EOR in carbonate reservoirs

Use of supercritical CO_2 for EOR stimulates oil production from low recovery oilfields, simultaneously contributing to minimizing the impact of CO_2-emission to the atmosphere. The injected CO_2 remains trapped in the underground geological formations, as much of the CO_2 is replacing the oil and water in the pores (NRG Energy, 2014).

Some of the world's largest remaining oil reserves are found in oil-wet, fractured carbonate reservoirs. The oil production performance from these reservoirs is nearly half the production from other reservoirs, whereas the CO_2 utilization is about 60% less (Ettehadtavakkol *et al*, 2014; Fitch, 2010). CO_2-EOR in carbonate reservoirs poses great challenges to the oil industry as it is strongly linked to the relationship between the fractures and the rock matrix. Because fractures may exhibit permeabilities that are several orders of magnitude higher than the permeability of the rock matrix, the CO_2 may channel into the high permeable fractures and thereby not contribute to EOR.

4 Simulations

The near-well simulations of CO_2-injection into the carbonate reservoir were carried out using the commercial reservoir simulation software Rocx, in combination with OLGA. The OLGA software is the main program, but several additional modules are developed to solve specific cases. The geometry for the simulated reservoir is 0.5 m in length, 96 m in width and 50 m in height. 3 grid blocks are defined in x-direction, 25 in y-direction and 10 in z-direction. The radius of the wellbore is 0.15 m. The well is located 35 m from the bottom, indicated as a black dot in Figure 4.

The reservoir is divided into three zones in x-direction. A constant porosity of 0.15 is used in the entire reservoir. A permeability of 4000 mD is set in the second zone, and a permeability of 40 mD is set in the first and the third zone. The second zone represents the fractured part, thus the permeability is set much higher in this zone compared to the two other zones. The temperature is maintained constant at 76°C and the waterdrive pressure from the bottom of the reservoir is 176 bar, the wellbore pressure is set to 130 bar.

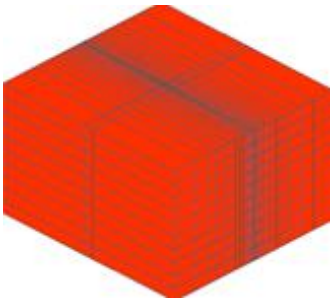


Figure 4. Grid and geometry of the simulated reservoir.

The reservoir and fluid properties for the simulations carried out are presented in Table 2.

Table 2. Reservoir and fluid properties for the specific simulations.

Properties	*Value*
Oil viscosity	10 cP
Reservoir pres sure	176 bar
Reservoir temperature	76°C
Oil specific gravity	0.8
Porosity	0.15
Permeability first zone (x-y-z direction)	40-40-20 mD
Permeability second zone (x-y-z direction)	4000-4000-2000 mD
Permeability third zone (x-y-z direction)	40-40-20 mD
Wellbore pressure	130 bar

The module Rocx is connected to OLGA by the near-well source component in OLGA, which allows importing the file created by Rocx. In order to get a simulation of the complex system including valves and packers, OLGA requires both a "Flowpath" and a "Pipeline" as shown in Figure 5.

In the simulations, the "Flowpath" represents the pipe and the "Pipeline" represents the annulus. The annulus is the space between the pipe and the rock, as presented in (Schlumberger, 2007).

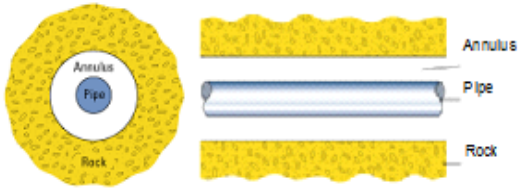


Figure 5. A schematic of the pipe and the annulus. (Schlumberger, 2007).

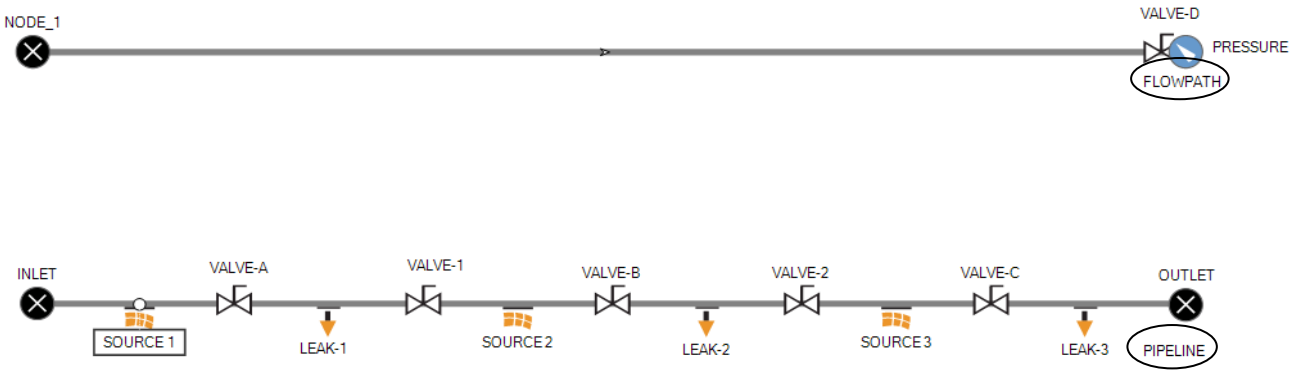


Figure 6. OLGA study case for the performed simulations.

Figure 7 illustrates how the "Flowpath" is divided into six equal sections. The sources implemented in the "Pipeline" are connected to the boundaries in Rocx, and indicate the inflow from the reservoir into the annulus. the leaks indicate the inflow from the annulus into the pipe, through the control valves A, B and C. The packers are simulated as closed valves and are installed to isolate the different production zones in the well.

In the simulations, the packers divide the "Pipeline" into three zones. The inflow from Source-1 goes from section one in the annulus and enters the pipe in section two. Similarly, for the flow in the production zones two and three.

4.1 Relative permeability curves

The simulation software Rocx generates the relative permeability curves for oil (K_{ro}) and water (K_{rw}). The calculations are based on the Corey correlation, a power law relationship with respect to water saturation. The model is derived from capillary pressure data and is widely accepted as a good approximation for relative permeability curves in a two-phase flow. The required input data is limited to the irreducible water saturation (S_{wc}) and the residual oil saturation (S_{or}), and their corresponding relative permeabilities (Tangen, 2012):

$$K_{rw} = K_{rwoc}\left(\frac{S_w - S_{wc}}{1 - S_{wc} - S_{or}}\right)^{n_w} \quad (1)$$

$$K_{ro} = K_{rowc}\left(\frac{1 - S_w - S_{or}}{1 - S_{wc} - S_{or}}\right)^{n_{ow}} \quad (2)$$

Where S_{wc} defines the maximum water saturation that a reservoir can retain without producing water, and S_{or} refers to the minimum oil saturation at which oil can be recovered by primary and secondary oil recovery. K_{rwoc} is the relative permeability of the water at the residual oil saturation, and K_{rowc} is the relative permeability of oil at the irreducible water saturation. n_w and n_{ow} are the Corey coefficients for water and oil respectively. The coefficients are functions of the pore size distribution in the reservoir and are therefore reservoir specific. The Corey coefficients strongly influence the relative permeability curves, as the relative permeability changes when the pore-geometry change. Typical values for the Corey coefficient for an oil-wet reservoir are n_w = 2-3 and n_{ow} = 6-8.

To simulate CO_2-injection into the reservoir, it was necessary to correlate for the effects of CO_2 to the

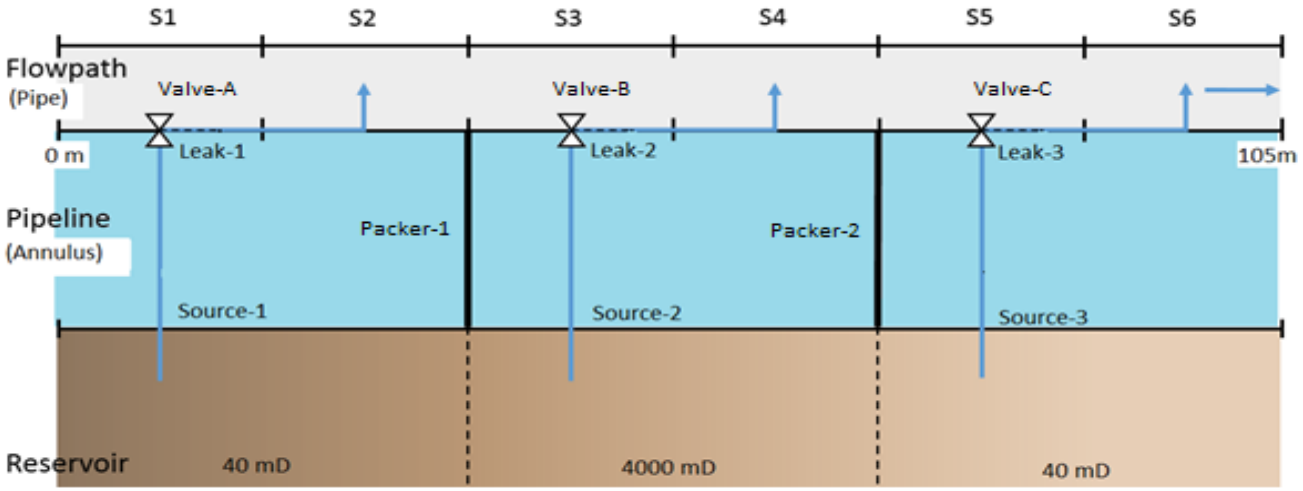


Figure 7. The near-well simulation in OLGA.

relative permeability curves. CO_2-injection reduces the interfacial tension and the oil viscosity, and causes oil swelling. Based on these parameters, the following relations was implemented to calculate the Corey's exponents and the residual oil saturation for simulation with CO_2-injection (Ghoodjani *et al*, 2011; Tangen, 2012):

$$n_{ow}(CO_2 - injection) = 0.568951 \cdot n_{ow} \quad (3)$$

$$S_{or}(CO_2 - injection) = 0.754288 \cdot S_{or} \quad (4)$$

The relative permeability curves for the performed simulations are generated in Rocx, using the parameters listed in Table 3.

Table 3. Relative permeability data for the specific simulations.

S_{wc}	S_{or}	K_{rowc}	K_{rwoc}	n_w	n_{ow}
0.1	0.1	1	0.75	3	3.4

Figure 8 shows the implemented relative permeability curves for the simulations. The green lines represent the relative permeability of oil (K_{ro}) and the blue lines represent the relative permeability of water (K_{rw}). Initially, when the water saturation is equal to the critical water saturation ($S_w = S_{wc}$) the injected CO_2 does not contact fully with the oil. As oil saturation decreases, the movement of oil becomes more difficult and the injected CO_2 will improve the oil flow by lowering interfacial tension and the oil viscosity. The CO_2-injection may lead to reduced trapped oil and lower residual oil saturation by swelling mechanism.

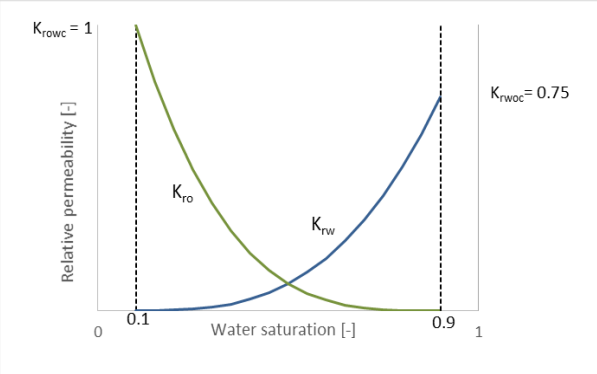


Figure 8. Relative permeability curves.

4.2 Input to OLGA and Rocx

The simulations were carried out for an oil-wet carbonate reservoir with fracture. Two different cases were simulated. Both cases include the relative permeability curves seen in Figure 8 and the reservoir and the fluid properties listed in Table 2. In Case 1, all control valves A, B and C are fully open. In Case 2 the control valve A and C are open, while control valve B is kept closed. This is to study how closing the fractured zone will affect the oil production and the CO_2-recycle ratio in the reservoir. The simulations where run for 400 days. Detailed specifications for the simulations are listed in Table 4.

Table 4. Input for the performed simulations.

Case	*Data input to Rocx*	*Relative perm. curve*	*CO2-injection*	*Position Valve A and Valve C*	*Position Valve B*
1	See Table 2	See Figure 9	Yes	Open	Open
2	See Table 2	See Figure 9	Yes	Open	Closed

5 Results

Produced water is the largest by-product associated with the oil production. The oil industry aim for new inflow control technology to shut off production from highly fractured zones when water breakthrough occur, and thereby be able to utilize the benefits from CO_2-EOR. To simulate the closing of the fractured zone in Case 2, Autonomous Inflow Control Valves (AICV) replace the conventional Inflow Control Devices (ICD) in the well. The AICV completely stops the production from a specific production zone when it starts to produce water and/or gas along with the oil. The oil production from the well will continue from the other production zones in the reservoir. Figure 8 and Figure 9 shows the accumulated oil and water production respectively. The orange lines represent Case 1 and the black lines represent Case 2. Both cases simulate injection of CO_2 into the reservoir, thus simulates CO_2 displacing the oil. Assuming that the injected CO_2 completely dissolve in the water phase, water is considered as carbonated water in the following results. The consequence with injection of CO_2 into fractured reservoirs is that the carbonated water and CO_2 moves through the fractures and directly into the production well, without being distributed in the reservoir. Large amounts of CO_2 will in that case be recycled and will not contribute to EOR.

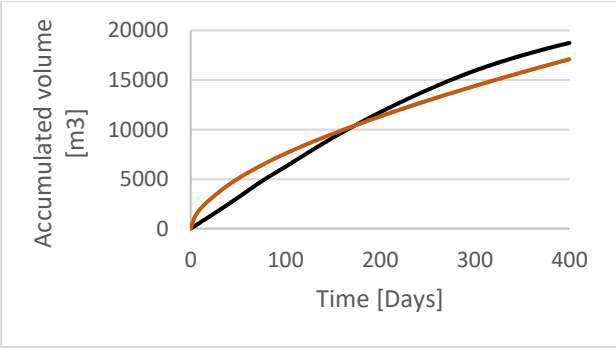


Figure 8. Accumulated oil volume.

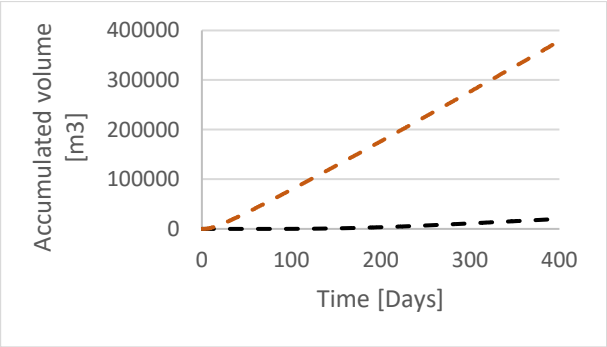


Figure 9. Accumulated water volume.

The large volume flow of carbonated water produced in Case 1 is due to no restrictions for the fluids to flow through the fractured zone in the reservoir, consequently most of the oil and carbonated water are produced from the second zone in this case. This is seen more clearly in Figure 10, where the total liquid flowrate in the different sections in the pipe is displayed. The total liquid flow includes the volume of oil, water and CO_2. The orange line represents Case 1 and the black line represents Case 2.

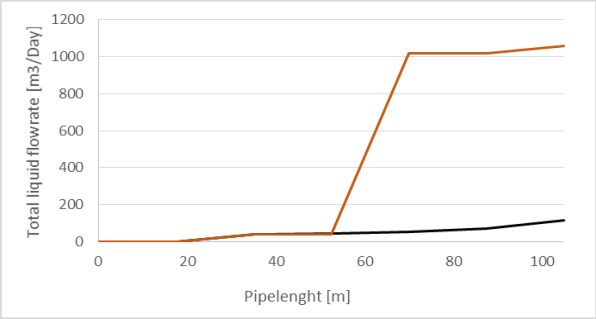


Figure 10. Total liquid flowrate along the pipe.

As seen from the figure, the major part of the total liquid volume is produced in the fractured part of the reservoir in Case 1, as it enters the pipe in section 4. For Case 2, the total liquid flowrate increases along with the pipeline, but the inflow from the annulus to the pipe occurs mainly in section 2 and section 6. This is as expected since these sections represent the first and the third production zone in the reservoir. In section 1, 3, 4 and 5 the total liquid flowrate is constant, this is due to no inflow from the annulus into the pipe in these sections, as the control valve in the second production zone is closed.

Fractured carbonate reservoirs are a major challenge for the oil industry using CO_2-EOR. The low permeability and the high heterogeneity result in high production rate of water and CO_2, mainly through the fractured zone. By chocking the high permeability zone, the oil and water production is reduced. On the second hand it delays the water breakthrough, which in turn results in decreased water cut. The water cut is the defined as the ratio of water to the total liquid, and expresses the amount of water produced along with the oil. Figure 11 shows the water cut during the whole simulation, the orange line represent Case 1 and the black line represent Case 2.

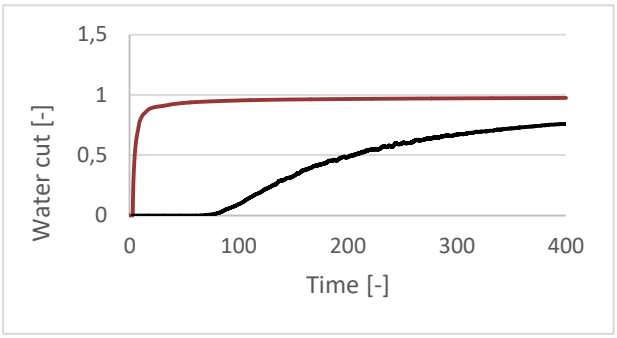


Figure 11. Water cut.

Carbonated water and CO_2 produced along with the oil cause economic, operational and environmental difficulties for the oil industry. Treatment of the produced water faces challenges resulting in necessity for expanding the capacity of water separation and facilities for handling the large volumes of carbonated water and CO_2. The high heterogeneity of carbonates is the main reason for the low oil recovery from these reservoirs. Water flows more easily through the fractured zone compared to the oil, resulting in very early water breakthrough and thereby high water and low oil flowrate. The results from the performed simulations are summarized in Table 5.

Table 5. Results from Case 1 and Case 2.

Case	*Accumulated oil volume [m3]*	*Accumulated water volume [m3]*	*Water breakthrough [day]*	*Water cut [-]*
1	17 000	355 500	2.9	0.975
2	18 700	20 100	64	0.725

As seen, closing the fractured zone in the reservoir is advantageous to achieve improved oil quality as well as lower production and separation costs. The water cut in Case 1 is 0.975 after 400 days of simulation, meaning that 97.5 % of the total liquid volume is carbonated water and CO_2. The high value for the water cut results in low quality oil and large amounts of carbonated water and CO_2 to be removed subsequently. Another problem with fractures is decreased recovery due to early breakthrough of carbonated water in the well. When this occurs the carbonated water will be the dominating fluid and reduce the production of oil. At a certain point the production will no longer be profitable and the well may be abandoned even though a large amount of oil is still left in the reservoir.

6 Conclusions

The objective of this work was to study CO_2-EOR in a fractured carbonate reservoir. The study included near-well simulations of oil production, using the reservoir software Rocx in combination with OLGA. CO_2-EOR in fractured carbonate reservoirs gives low oil production, high water production, early water breakthrough and high water cut. The enormous amounts of water produced result in challenges regarding the necessity for expanding the capacity of water separation and facilities for handling the large volumes of carbonated water and CO_2.

Fractures in the reservoir are a major problem for the oil industry using CO_2-EOR. Water breakthrough occurs after only 2.9 days in a fractured reservoir and the water cut is 97.5 % after 400 days of production. Due to the very early breakthrough of water, significant amounts of the injected CO_2 will be recycled with the produced water. The simulations indicate that CO_2-injection into a carbonate reservoir in combination with closing the fractured zone causes delayed water breakthrough and dramatically reduced water cut, resulting in improved oil quality, longer lifetime for the well as well as lower production and separation costs.

References

K. Ahmed. *CO_2 injeksjon for økt oljeutvinning i kalk.* Masteroppgave i reservoarfysikk, Universitet i Bergen: Institutt for fysikk og teknologi, 2013.

M. Brettvik. *Experimental and computational study of CO_2 for EOR and secure storage reservoirs*. Master Thesis, Telemark University College, Faculty of Technology, 2013.

M. R. Esfahani and M. Haghigi. Wettability evaluation of Iranian carbonate formations. *Journal of Petroleum Science Engineering*, 92(2-4): 257-265, 2004.

A. Ettehadtavakkol, L. W. Lake and S. L Bryant. CO_2-EOR and storage design optimization. *International Journal of Greenhouse Gas Control*, 25: 79-92, 2014.

P. J. R. Fitch. (2010). *Heterogeneity in the petrophysical properties of carbonate reservoirs.* Doctor of philosophy, The University of Leicester, Department of Geology, 2010.

E. Ghoodjani and S. H. Bolouri. *Experimental study and calculation of CO_2-oil relative permeability*, 53-(2): 123-131, 2011. Iran: Petroleum & Coal Sharif University of Technology and Shahid Bahonar University. ISSN 1337-7027.

Å. Haugen. *Fluid Flow in Fractured Carbonates: Wettability Effects and Enhanced Oil Recovery*. PhD-dissertation, Department of Physics and Technology, University of Bergen, Norway, 2010.

B. Hill, S. Hovorka and S. Melzer. Geologic carbon storage through enhanced oil recovery. *Energy Procedia*, 37: 6808-6830, 2013. USA, Elsevier Ltd.

International Human Resources Development Corporation. IPIMS, e-learning for the Upstream Petroleum Industry, Irreducible water: www.ihrdc.com/els/ipims-demo/t26/offline_IPIMS_s23560/resources/data/G4108.htm, 02.02.2016.

C. H. Moore. *Carbonate Diagenesis and Porosity*, 46, USA, Elsevier Science Publishers B.V. 1989. ISBN: 0-444-87415-1.

NRG Energy. CO_2 Enhanced Oil Recovery. *NRG Fact Sheet*, Texas: NRG Energy Inc., Available from: http://www.nrg.com/documents/business/pla-2014-eor.pdf, 2014.

Oil and Gas 360. Enhanced Oil Recovery Can Reduce World's Carbon Emissions: Western Governors, Available from:http://www.oilandgas360.com/enhanced-oil-recovery-can-reduce-worlds-carbon-emissions-western-governors, 02.02.2016.

S. M. Pasala. *CO_2 displacement mechanisms: Phase equilibria effects and Carbon dioxide sequestration studies*. Doctor of Philosophy, University of Utah: Department of Chemical Engineering, 2010.

A. Satter, G. M. Iqbal and J. L. Buchwalter. Practical Enhanced Reservoir Engineering, Assisted with Simulation Software. USA, PennWell Corporation. ISBN-13: 987-1-59370-056-0. ISBN-10: 1-59370-056-3, 2007.

Schlumberger. Fundamentals of Wettability. *Oilfield Review*, 2(19): 66-71, 2007.

M. Tangen. *Wettability Variations within the North Sea Oil Field Frøy*. Master Thesis, Norwegian University of Science and Technology, Department of Petroleum Engineering and Applied Geophysics, 2012.

T. A. Tarek. *Petrophysical characterization of the effect of Xanthan gum on drainage relative permeability characteristics using synthetic unconsolidated core plugs.* Master of Petroleum Engineering, Dalhousie University, Halifax, Faculty of Engineering, 2014.

Performance of Electrical Power Network with Variable Load Simulation

Ahmed Al Ameri and Cristian Nichita

Groupe de Recherche en Electrotechnique et Automatique, GREAH Lab., University of Le Havre, France,
ahmed.al-ameri@etu.univ-lehavre.fr, cristian.nichita@univ-lehavre.fr

Abstract

Today's system operators face the big challenge of constructing simulation of systems that make efficient select of generation resources under variable load profiles. This paper describes IEEE five bus system modeling which simulated under Simulink. The real power load model designed to allow different load profile types (residential, commercial and industrial) connecting to load buses. The main purpose of this paper is to demonstrate the performance of power network based on load profile modeling as a means for enhance Distribution Network Operators (DNOs) decision in power systems. In this paper IEEE five bus system is used as a test bed. The results are shown with constant and variable load model. The results indicate the effectiveness of this flexible load profile model applied to the five bus system.

Keywords: power system analysis, load modelling, simulation, load profile, network modelling

1 Introduction

Deregulation of the electric energy systems and a progressive increase of the load play a key role in improving reliability and continuity of the electrical utility services they provide. Due to the complexity of load profile, many challenges facing electrical power system analysis, like, state estimation, load flow calculation, and network planning and operation. Distribution network operators (DNOs) report the performance of their network based on inadequate modelling approaches because large power systems are simply represented by a bulk load.

The hardest part for most DNOs is to determine the errors between estimated and actual load profile simulation. There is some of early-published state estimation studies describing the role of load profiling on the classical electrical power networks. Author in (Jardini et al., 2000), presents a statistic analysis of different load curves to obtain the representative curves of the most important customers. The measurements were performed for residential, commercial and industrial to determine customers' daily load profile. The fuzzy c-means (FCM) method is used for load profiling in (Chang et al., 2003). The load profile assignment performed by using customers' monthly energy usage data in two steps. After load profiling, the author used recognition technique to assign 500 customers. In (Kim et al., 2011), It was interesting to used load profile for energy diagnosis system. Different load profile data according to customer type (residential, commercial and industrial) obtained from metering devices. Consumption separation method was used to classify duration of load devices to a certain time interval (1 hour, 1 day, 1 week and 1 month).

On other hand, other authors modelling the optimal power flow control for dynamic grid where the load changes according to a profile. (Almeida et al., 2000) presents an algorithm to calculate a sequence of Primal-Dual Interior Point optimization solutions under variable load conditions. This methodology to find optimal power flow based on two main steps: predictor, which estimate a new operating point for an increment in the load by linear approximation; and corrector step, which uses non-linear method to find the optimal solution to the new load level. A three load profile are used to find the optimal reactive power control. The daily load curves divided into several sequential levels to reduce operations of control devices switching. Voltage quality and power loss for different loads were consider as objective function for optimization techniques. Typical clustering method and heuristic iteration technique used the maximum load deviation (MLD) to decompose the load curve (Varga et al., 2015).

Other opportunities to improve planning and operation of modern network with/without distributed generation and storage under different load conditions are considered in a large number of articles recently (Hernando-Gil et al., 2016; Bazrafshan et al., 2017; Guggilam et al., 2016; Ma et al., 2016; Yi et al., 2016; Thrampoulidis et al., 2016; Geng et al., 2016, Li et al.,2016). This fact shows that this subject continues an interesting topic of research. In relation to modern network, a real-time classification and encoding of load profiles has been proposed in (Varga et al., 2015). The author presents software framework to manage the load profile at power system operation. The framework is based on artificial neural network as encoding engine and local hashing algorithm as classifier engine. A dynamic load profile was cluster and classify by multi-resolution analysis (MRA) method (Bazrafshan et al., 2017). The ability of traditional methods in profiling load developed by MRA method for three key (large,

volatility and uncertain smart metering data). A more flexible load profiling with less computation presented in this method by three main steps (decomposition, clustering and reconstruction).

Moreover, there has been an increasing amount of literature on the planning of modern power system's reliability. In active distribution systems, (Hernando-Gil et al., 2016) proposed methodology based on empirical load profile and time varying fault probabilities for reliability planning and risk estimation. The approach is developed to avoid the underestimation of network's performance at bulk supply points for more realistic estimation of customer interruptions.

Research has shown that installation of distributed generation and storage energy take more attention. Radial distribution networks with photovoltaic (PV) generation was tested to optimize the real power consumption of loads (Bazrafshan et al., 2017). The power management scheme developed to determine the optimal demand response schedule that accounts for variable real power injection by PV units. So that, the programmable loads provide opportunity to reduce the peak load in periods of inappropriate generation. In distribution feeders, load profile data and realistic photovoltaic (PV) generation are utilized to optimize its operation. The active and reactive power set points for PV was determine according to voltage regulation and a variety of objective functions (Guggilam et al., 2016). The proposed method leverage a linearized version to formulate a quadratic constrained quadratic program (QCQP) as direct applications to distribution networks with PV systems. The cost efficiency of the residential electricity consumption improved by load scheduling (Ma et al., 2016). The load scheduling framework based on fractional programing approach to develop a cost efficient for the demand side's day-ahead process and real-time pricing mechanism. The proposed algorithm considered the distributed energy resources and service free in their framework.

Finally, in attempt to improve the power system operation more effectively, energy storage systems installed (ESS) with/without distributed generation to the modern grid. Real load connecting to distribution networks has tested to schedule the ESS by Monte Carlo simulation (Yi et al., 2016). The optimization technique used for solving an ESS scheduling problem considering real load, variable wind energy sources and transmission line real time thermal rating (RTTR). The load shifting by optimize placement, sizing and control of energy storage system presented in (Thrampoulidis et al., 2016). The network topologies with regardless/regard of the load demand, generation capacities and line flow limitations effected the costs. A charging/discharging policy for the installed storage units formulated as slower time-scales.

However, investigating and modelling varying energy demands in various sectors (residential, commercial and industrial) will cause significant changes in planning, operation and control of power system. Most studies in load modelling and profiling into electrical power network have been carried out in a small number of area. This paper attempts to simulate the electrical network with different load profile, which can help DNOs to avert the underestimation of network's performance at bulk generation points for more actual estimation of customer interruptions. Load demand has been developed to determine the overall power flow and identify generation sources required to meet its increasing as well as exceeded rating elements in the network.

The rest of the paper is constructed as follows. Section 2 demonstrates the element of five node network modelling. Section 3 gives a brief overview of the load profile conception for different type of customers. Simulation is implemented and results are presented in Section 4. Conclusions are drawn in Section 5.

2 Load Profile

In power system, a load profile is a graph illustrating the variation in the demand/electrical load versus time. A load profile will vary according to temperature, holiday seasons and customer type (typical examples include residential, commercial and industrial),. DNOs use this information to plan how much electricity they will need to generate at any given period. These load curves are useful in the selection of generator units for providing electricity.

Direct metering devices such as smart grid meters, data logging sub-meters, utility meter load profilers and portable data loggers can determine load profiles. Real demand can be collected at strategic locations to analysis load performance, which is beneficial to both distribution and end-user customers looking for peak consumption (Geng et al., 2010). For most customers, based on meter reading schedules, consumption is measured on a monthly. Load profiles are used to convert the monthly consumption data into estimates of hourly or sub hourly consumption in order to determine the electrical utilities obligation. For each hour, these estimates are aggregated for all customers of an energy provider, and the aggregate amount is used as the total demand that must be covered by the utilities.

In this section, the detailed simulation of the load profile would be described. For brevity purpose, detailed simulation of the load profile for residential, commercial and industrial would be shown, covering the lower and higher side of the residential units. Simulation of three bus load are described concisely.

2.1 Residential Consumers

More than half of all electrical power is consumed by residences type, which vary in their daily activity patterns and the types of appliance they own. Load

varies by time of day and year where its curve shape be function of consumer demand. Figure 1 illustrating how electrical demand characteristics varies over a day, or when house owners are using electrical power.

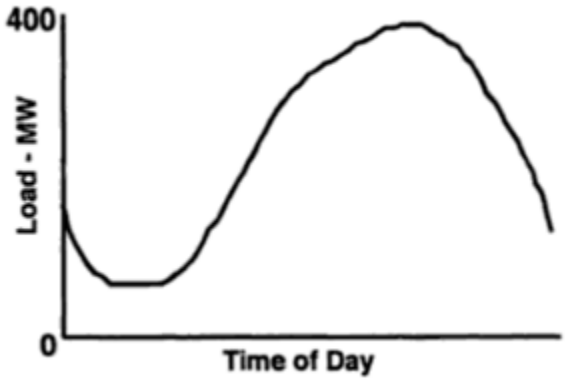


Figure 1. Example of residential consumption.

Due to the differences in electric appliances, the definition of the representative curves of a range is not an easy task to be done and the people habits rising to curves of shapes in the peak. In some countries, the residential energy consumption value is mainly due to refrigerator or freezer, whilst the water heating gives the curve peak where heater resistance take 8 min (Li et al., 2016). Therefore, it is very hard to characterize the peak demand because load pattern not fixed for all residential usage and depends on many factories such as weather, type of human work etc.

2.2 Commercial and Industrial Consumers

In commercial businesses load, small and large consumers having similar end uses to residential (cooling, heating etc) in addition to many need to commercial devices (office machinery, cash register, escalators etc). Figure 2 represents the electrical use of a commercial facility during 24 hours. The commercial load (office building, restaurants stores etc) shows a strong upward curve during summer (or winter) session because it depend heavily on cooling (or heating) systems (Lee, 2004).

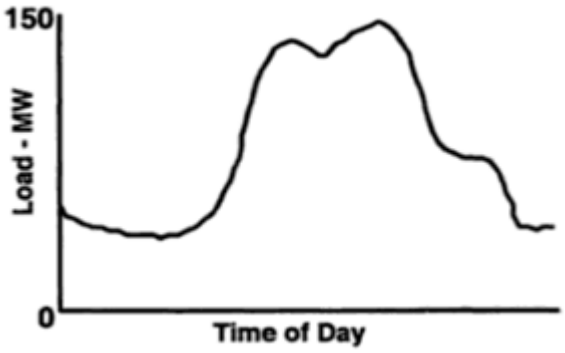


Figure 2. Example of commercial consumption.

Finally, industrial facilities and plants use electricity to variety of manufacturing applications such as compressor motors, heating systems etc. The industrial load profile does not vary as much through the day where it depend on work, weekends and break times. The peak demand of summer day for industrial consumption from utility system shown in Figure 3.

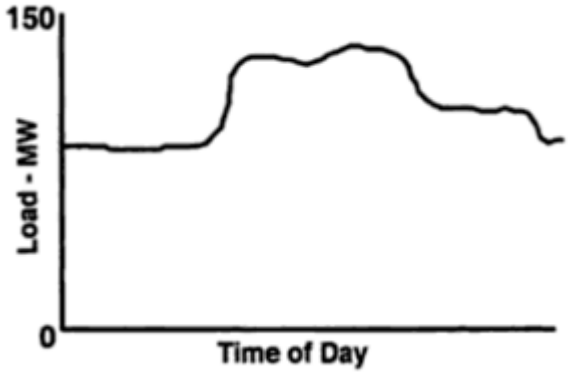


Figure 3. Example of industrial consumption.

3 5 Node Network Modelling

Throughout the electrical power network there are common buses that look like important branch points within the power grid. These buses operate at a defined voltage level and phase angle to forming the complex bus voltage. In general, three types of busses is consider in a power network, namely the slack bus, the generator bus and the load bus.

Figure 4 shows a single line diagram of a 5 bus system with two generating units, seven lines. Per unit system based on 100 MVA was considered for all parts of network. Four basic parts of the system are modelled: slack generator, PV control generator, transmission lines and load profiles. Figure 5 presents the simulation of the 5 bus network using Simulink.

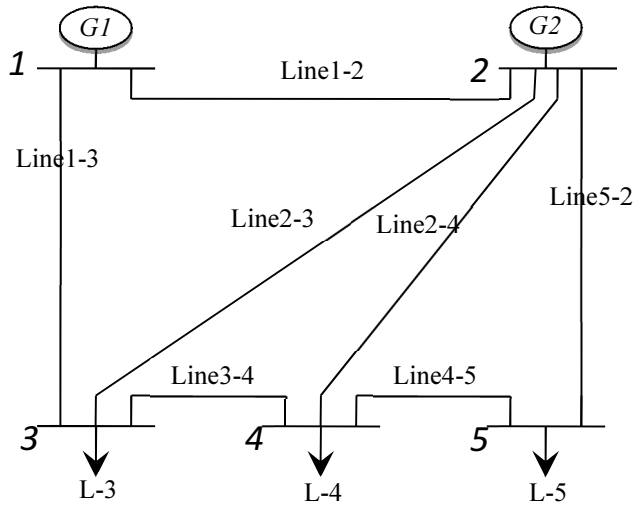


Figure 4. Single line diagram of 5 bus network.

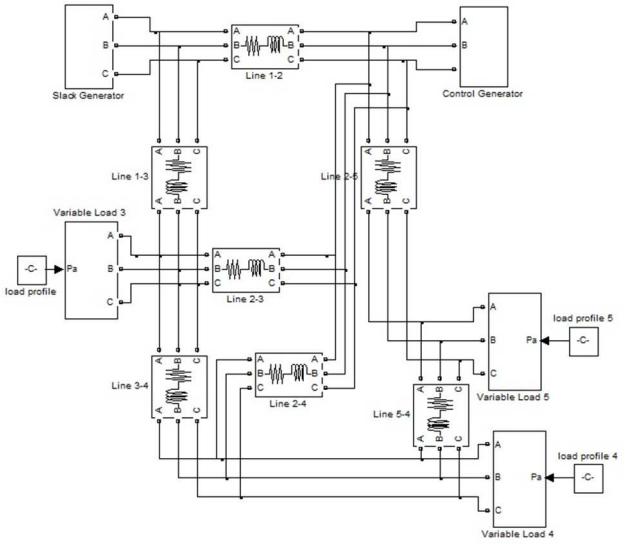


Figure 5. 5 bus network simulation.

All generation units, load demand and power loss calculated is shown in Table 1.

Table 1. Generation and load demand for IEEE 5bus.

Unit	*Total (MW)*
Total Generation (PG2)	unlimited
Total Generation (PG2)	90
Max. Demand(PD3) residential	45
Max. Demand(PD4) Commercial	40
Max. Demand(PD5) Industrial	60

3.1 Slack generator

In a simulated power system, some quantities allowed to vary or swing to solve particular steady-state problem successfully. For that, there is only one slack generator has known voltage magnitude |V| and voltage phase (set to $1.0\angle 0.0^{\circ}$ (per-unit)). In Figure 6, Psg and Qsg are the swing variables and obtained through the load flow solution as follows:

$$S_{sg} = P_{sg} + j\,Q_{sg} \qquad (1)$$

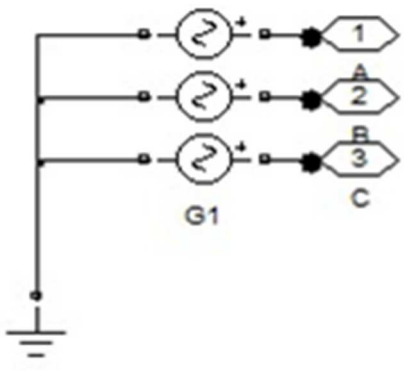


Figure 6. Slack generator simulator.

3.2 Voltage Control Generator

Generator buses or voltage controlled buses have inputs of the voltage magnitude corresponding to the generator voltage and real power Pg corresponds to its rating (Figure 7) . Generally, voltage controlled busses are connected to equipment used for voltage and VAR correction, such as static VAR correction systems, generators and shunt capacitors (Figure 8). It is calculated the reactive power generation Qg and phase angle of the bus voltage by load flow solution. In general, generator is modeled as a complex power injection at a specific bus (i) is

$$s_g^i = p_g^i + jq_g^i \qquad (2)$$

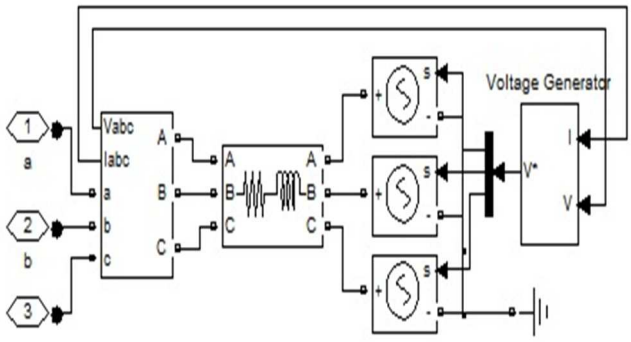


Figure 7. Voltage control generator.

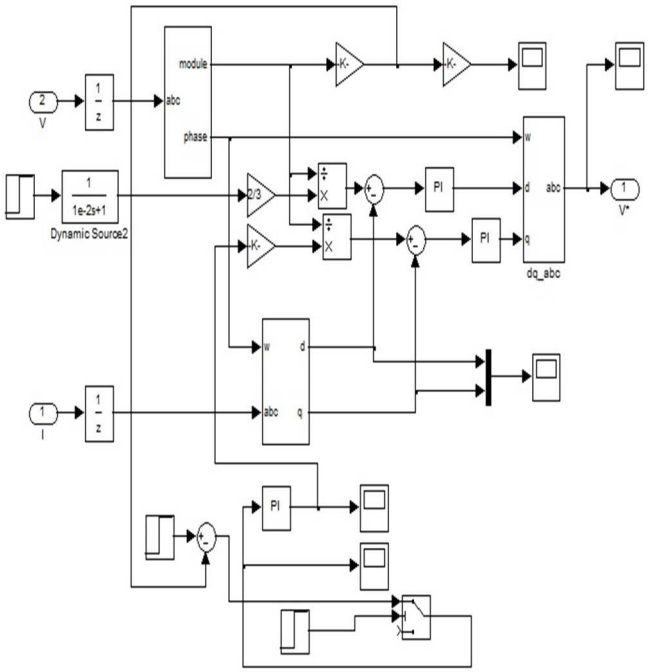


Figure 8. Voltage regulator.

3.3 Transmission Line

This section deals with the modelling of transmission line elements encountered in the electrical power network. Transmission line transmits bulk power from sending to receiving end and represented by standard (π model) consist of four main elements (resistance, inductance, capacitance and conductance). The analysis of power system is mainly dependent on the performance of the transmission line in the power grid. All transmission lines, transformers, and phase shifters are modeled with a common branch model as shown in Figure 9. The three phase series RLC branch block (Figure 10) used to simulate performance of transmission line by setting its parameter corresponding data of network (Milano, 2005).

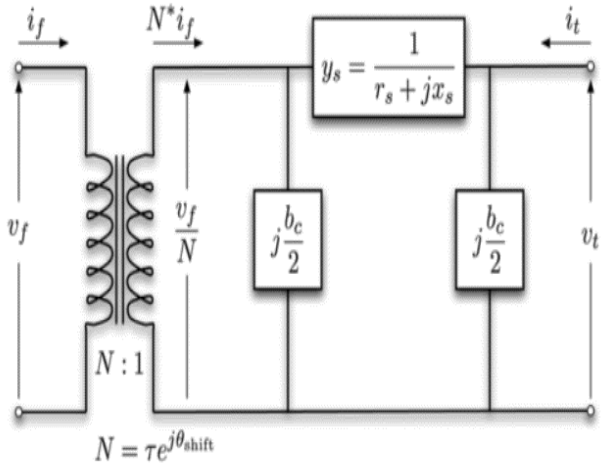


Figure 9. π model.

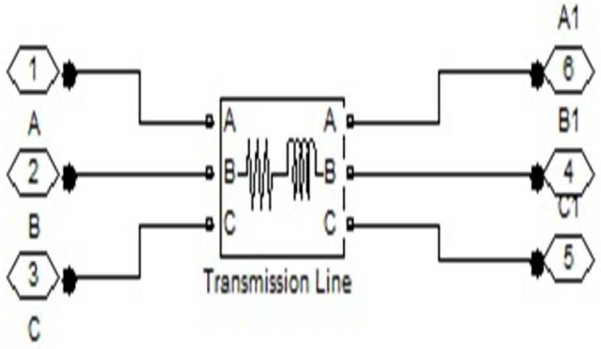


Figure 10. Transmission line simulation.

3.4 Load bus

At this bus, the real (Pd) and reactive power load (Qd) are specified and no generator is connected to it. It is required to find out the voltage magnitude and phase angle through load flow calculations. The total demand of the system is represent the distributed loads over the whole network. Power Systems on whether they represent an industrial, commercial or residential load, they can vary greatly in electrical characteristics as well as quantity. In most power simulation, it sees the load as a simplest PQ load characteristic with constant demand of real and reactive power that does not change with any external influences. Various load models have been introduced, taking into account day, month and year cycles as well as it can be consider the voltage and frequency dependencies in the future work. Normally constant power loads are modeled as real and reactive power consumed at a bus (i) as follows:

$$\mathbf{s_d^i = p_d^i + jq_d^i} \quad (1)$$

The loads can be modeled using Simulink block (three phase series RLC load) but it will be represented as fixed (static) load. A constant MVA load model have no ability to vary with time. Therefore, that, simulation for variable real power load construct with fixed bus voltage reference and variable real power as shown in Figure 19. Fault or any external changes of network state will not effect on load parameters. This load models can be described by the following equation:

$$\boldsymbol{P_d = P_o\left(\frac{V}{V_o}\right)} \quad (2)$$

$$\boldsymbol{Q_d = Q_o\left(\frac{V}{V_o}\right)} \quad (3)$$

where $\boldsymbol{Q_o}$ stand for reactive powers consumed at reference voltage $\boldsymbol{V_o}$ and represented by three phase series RLC load while $\boldsymbol{P_o}$is vary according to load profile curve connected to this model. The external control structure block connected to this model can be variable load curve (continues or discrete). This model will enhance the ability of the system to be studied at different loading conditions.

4 Simulation Results

Before modelling the IEEE 5 bus, Newton Raphson method has been implemented to calculate all the parameters of the system. All generation units, load demand and power loss calculated is shown in Table 2.

Table 2. Generation, load demand and losses for IEEE 5bus.

Unit	*Total (MW)*
Total Generation (PG1,PG2)	148.05
Total Demand (PD3, PD4, PD5)	145
Total real power loss	3.05

The 5 bus IEEE modelling has been developed with Simulink in order to study its behavior under different load conditions. Slack generator simulated to has unlimited real and reactive power generation while its voltage and voltage angel set to 1.06∠0.0˚ per unit. Parameters of generation, load, voltage and voltage angel for other buses have set according to data of IEEE five bus shown in Table 3.

Table 3. IEEE 5 Bus Data.

Bus No.	*Type*	*Generation*		*Load*		*Voltage*	
		Real	*React.*	*Real*	*React.*	*Magn.*	*angle*
1	1	0	0	0	0	1.06	0
2	2	90	0	0	0	1	0
3	0	0	0	45	15	1	0
4	0	0	0	40	5	1	0
5	0	0	0	60	10	1	0

At the beginning, load is considered as constant and simulation implemented for 10 seconds to compare with Newton Raphson (NR) method results. Figure 11 and Figure 12 shows measurements of real power flow (RPF) and losses (RPL), respectively, in all the transmission lines. This measurements has been compared with accurate calculated by Newton Raphson method to valid Simulink model as in Table 4.

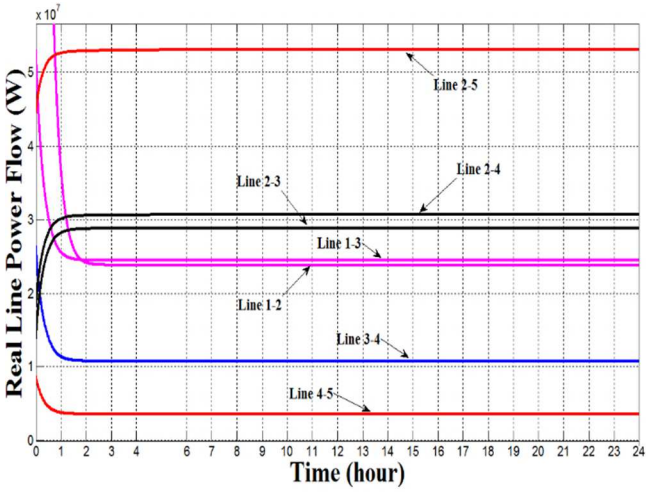


Figure 11. Active power flow on the lines with constant load.

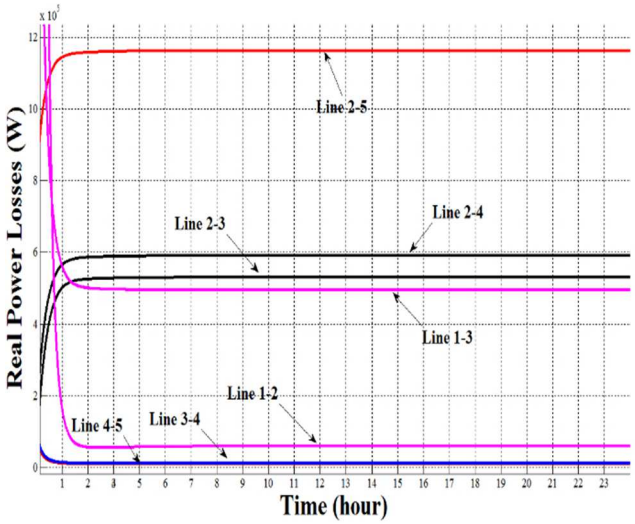


Figure 12. Active power losses on the lines with constant load.

In order to validate the proposed variable load simulation, random load profile used as input signal to Figure 19 instead of constant one. In Figure 13, output and output signals are identical except very small transient when signal switch from value to another.

Table 4. Real power lines flow and losses by NR.

	Transmission Lines						
	1-2	*1-3*	*2-3*	*2-4*	*2-5*	*3-4*	*5-4*
RPF (MW)	15.09	27.05	30.50	32.60	56.06	11.76	3.97
RPL (MW)	0.069	0.55	0.525	0.594	0.429	0.013	0.013

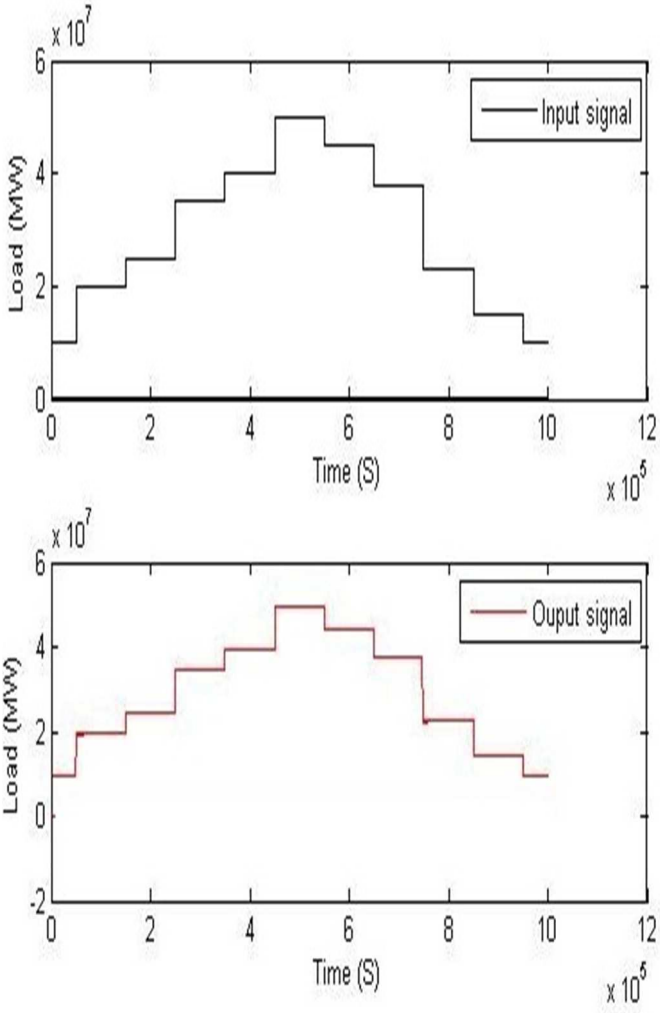


Figure 13. Load model testing.

Three different daily load profile (residential, commercial and industrial) are connected to load buses as follows: residential load curve (Figure 14) connect to bus 3, commercial load curve (Figure 15) connect to bus 4 and industrial load curve (Figure 16) connect to bus 5.

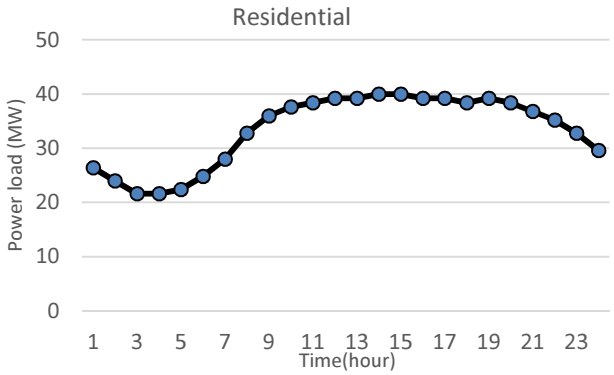


Figure 14. Residential daily load curve connect to bus 3.

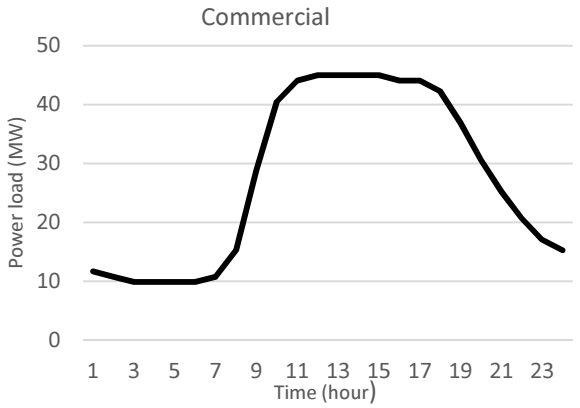


Figure 15. Commercial daily load curve connect to bus 4.

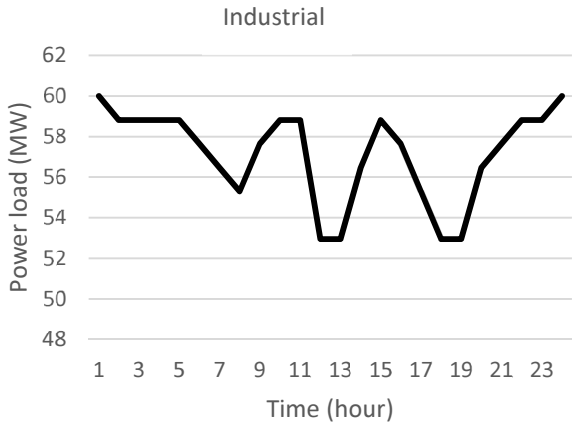


Figure 16. Industrial daily load curve connect to bus 5.

The results show that real power flow and losses in transmission lines will be effect by variable load profiles as shown in Figure 17 and Figure 18. The simulation of variable load details in Figure 19. The 5 bus Simulink simulation under different load conditions is performed on a computer with core i74800MQ CPU, 2.7GHz, 16GB RAM and Microsoft Windows 7 operating system running in real-time mode.

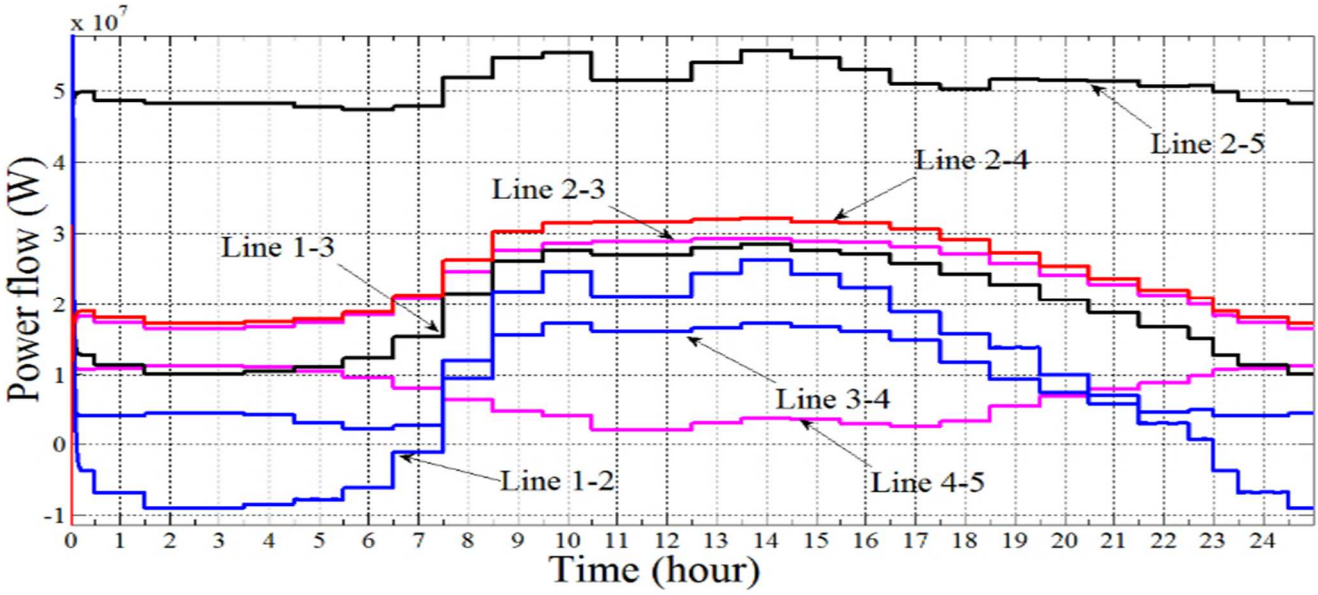


Figure 17. Lines Power flow with variable loads.

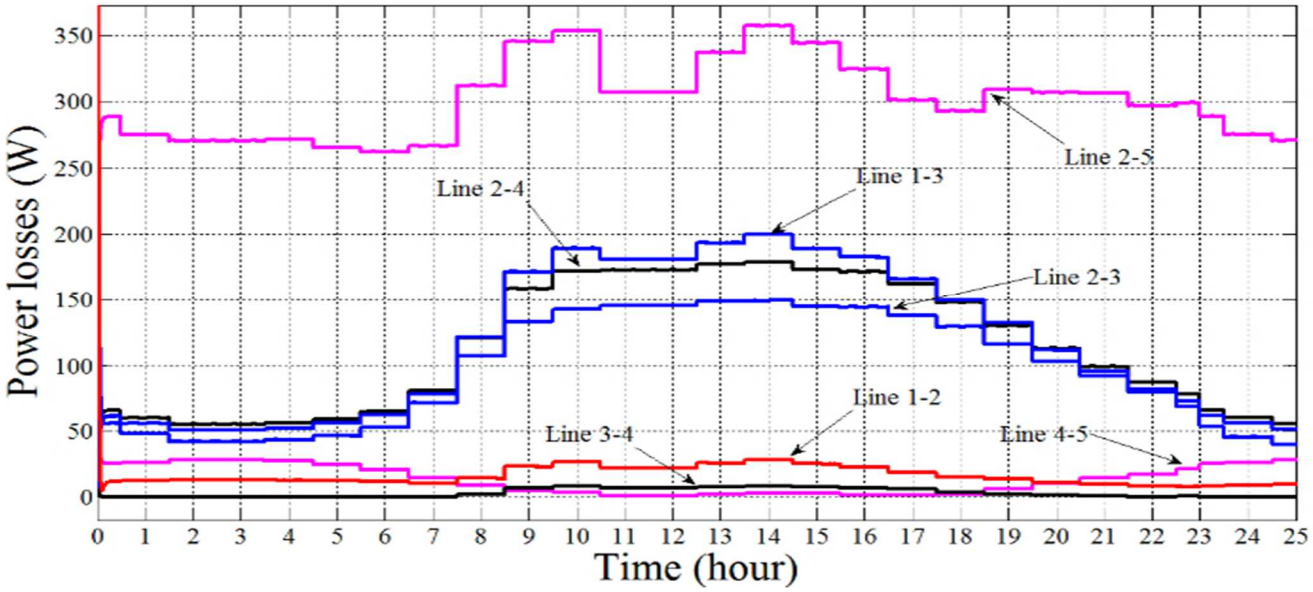


Figure 18. Lines losses with variable loads.

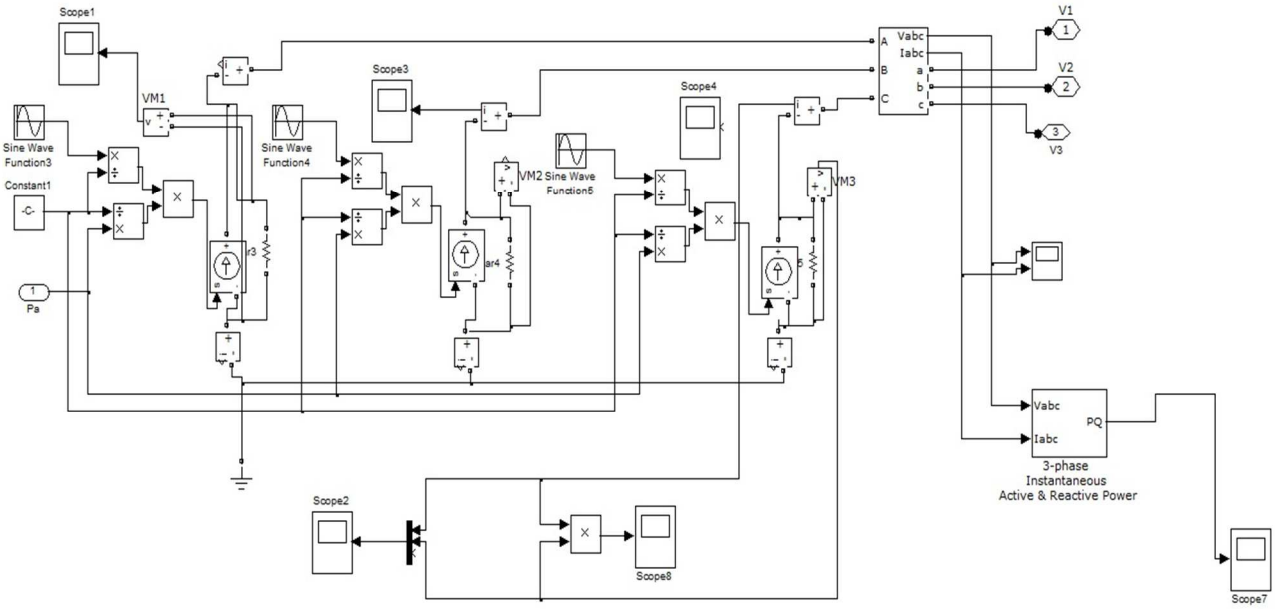


Figure 19. Variable load simulation.

5 Conclusions

This paper has presented network modelling with variable load profile using Simulink. . System of IEEE five bus is used as a test bed. This simulation can be easily adapted to accommodate different types of generation resources by considering time-varying load profile. The simulation introduce four main blocks represent slack generator, control generator, transmission line and variable load. Three different types of load profile (residential, commercial and industrial) tested the five bus network simulation. The results demonstrate that the simulation can be adequate to identify the real power flow and losses into transmission lines. Simulation results indicate that varying demand can change the dynamic performance of the system and will help DNOs understand what they need to do to provide solutions for network stability.

References

Katia C. Almeida and Roberto Salgado. Optimal Power Flow Solutions Under Variable Load Conditions. *IEEE Transactions on Power systems*, 15(4): 1204 - 1211, 2000.

Mohammadhafez Bazrafshan and Nikolaos Gatsis. Decentralized Stochastic Optimal Power Flow in Radial Networks With Distributed Generation. *IEEE Transaction on Smart Grid*, 8(2): 787 - 801, 2017.

R. F. Chang and C. N. Lu. Load profiling and its applications in power market. *In IEEE Power Engineering Society General Meeting,* Toronto, 2003.

G. Geng, J Liang, R. G. Harley, and R. Qu. Load Profile Partitioning and Dynamic Reactive Power Optimization International. *In Conference on Power System Technology (POWERCON)*, pp. 1-8, Hangzhou, 2010.

Swaroop S. Guggilam, Emiliano Dall'Anese, Yu Christine Chen, Sairaj V. Dhople, and Georgios B. Giannakis. Scalable Optimization Methods for Distribution Networks with High PV Integration. *IEEE Transactions on Smart Grid*, 7(4): 2061-2070, 2016.

Ignacio Hernando-Gil, Irinel-Sorin Ilie, and Sasa Z. Djokic. Reliability planning of active distribution systems incorporating regulator requirements and network-reliability equivalents. *IET Generation, Transmission & Distribution*, 10(1): 93 – 106, 2016.

J.A. Jardini, C.M.V. Tahan, M.R. Gouvea, S.U. Ahn, and F.M. Figueiredo. Daily Load Profiles for Residential, Commercial and Industrial Low Voltage Consumers. *IEEE Transactions on Power Delivery*, 15(1):375 - 380, 2000.

Sun Ic Kim, Hae Soon Kim, Yong Jae Joo, and Ji Hyun Kim. Power usage pattern and consumption separation method by load devices based on remote metering system's Load profile data. *In 11th International Conference on Control, Automation and Systems (ICCAS)*, Gyeonggi-do, 2011.

Willis H. Lee. *Power Distribution Planning Reference Book, Second Edition, Revised and Expanded*. New York : Marcel Dekker, Inc., 2004.

Ran Li, Furong Li, and Nathan D. Smith. Multi-Resolution Load Profile Clustering for Smart Metering Data. *IEEE Transactions on Power Systems*, 31(6): 4473-4482, 2016.

Jinghuan Ma, He Henry Chen, Lingyang Song, and Yonghui Li. Residential Load Scheduling in Smart Grid: A Cost Efficiency Perspective. *IEEE Transaction on Smart Grid*, 7(2): 771 - 784, 2016.

F. Milano. An open source power system analysis toolbox. *IEEE Transactions on Power systems*, 20(3): 1199–1206, 2016.

Christos Thrampoulidis, Subhonmesh Bose, and Babak Hassibi. Optimal Placement of Distributed Energy Storage in Power Networks. *IEEE Transactions on Automatic Control*, 61(2): 416 - 429, 2016.

Ervin D. Varga, Christian Noce, and Gianluca Sapienza. Robust Real-Time Load Profile Encoding and Classification Framework for Efficient Power Systems Operation. *IEEE Transaction on Power Systems*, 30(4): 1897 - 1904, 2015.

Jialiang Yi, Pádraig F. Lyons, Peter J. Davison, Pengfei Wang, and Philip C. Taylor. Robust Scheduling Scheme for Energy Storage to Facilitate High Penetration of Renewables. *IEEE Transaction on Sustainable Energy*, 7(2):797 - 807, 2016.

Simulation of CO_2 for Enhanced Oil Recovery

Ludmila Vesjolaja Ambrose Ugwu Arash Abbasi Emmanuel Okoye
Britt M. E. Moldestad

Department of Process, Energy and Environmental Technology, University College of Southeast Norway, Norway,
britt.moldestad@usn.no

Abstract

CO_2-EOR is one of the main methods for tertiary oil recovery. The injection of CO_2 does not only improve oil recovery, but also contribute to the mitigation of greenhouse gas emissions. In this study, near well simulations were performed to determine the optimum differential pressure and evaluate the effect of CO_2 injection in oil recovery. By varying the drawdown from 3 bar to 20 bar, the most suitable differential pressure for the simulations was found to be 10 bar. The effect of CO_2 injection on oil recovery was simulated by adjusting the relative permeability curves using Corey and STONE II correlations. By decreasing the residual oil saturation from 0.3 to 0.15 due to CO_2 injection, the oil recovery factor increased from 0.52 to 0.59 and the water production decreased by 22%.

Keywords: CO_2, enhanced oil recovery, relative permeability, near well simulation

1 Introduction

According to Melzer (2012) and Jelmert *et al* (2010), after water flooding CO_2-EOR is the most commonly used method for improved oil recovery. CO_2-EOR can be used for increased oil production in combination with CO_2 storage to mitigate CO_2 emissions. The method is widely used in the United States where the price of CO_2 is relatively low due to large resources of natural CO_2. More than 70 operating fields in USA use CO_2 for enhanced oil recovery. The majority of oil fields use closed-loop systems during CO_2-EOR. The working principle of CO_2-EOR is depicted in Figure 1. CO_2 is injected into the reservoir using injection wells. When CO_2 comes in contact with oil in the reservoir, the oil properties change and the oil becomes more mobile. In addition, the injected CO_2 displaces the oil and forces it to move towards the production well. Significant amount of the injected CO_2 is retained inside the reservoir pores and some amounts are produced together with the oil to the surface. On the surface, CO_2 is separated from the oil and re-injected into the reservoir, and is thereby giving rise to a closed loop system. The CO_2 which is separated from the oil, can also be injected to the underlying aquifer for sequestration (Jelmert *et al*, 2010; Zhang *et al*, 2015).

1.1 Mechanism of CO2-EOR

Crude oil contains hundreds of hydrocarbons and many of them contain more than 30 carbon atoms. CO_2 is miscible in hydrocarbons with less than 13 carbon atoms. CO_2 becomes mutually soluble with the immobile oil as the light hydrocarbons from the oil dissolves in the CO_2 and CO_2 dissolves in the oil. When CO_2 and oil are miscible, the interfacial tension disappears. This means that the physical forces holding the two phases apart are no longer present, which make it possible for the CO_2 to displace the oil that is trapped in the pores of the rock. The efficiency of CO_2-EOR is dependent on the miscibility of CO_2 in oil, and the miscibility is strongly affected by pressure. The solubility of CO_2 in oil depends on the type of oil where higher amount of CO_2 can be solved in light oil than in heavy oil. Mobility of the oil increases mainly due to interfacial tension reduction, oil viscosity reduction, oil swelling and due to acid effect on rock (Ghoodjani and Bolouri, 2011).

Interfacial tension strongly influences relative permeability between CO_2 and oil. Due to dissolution inside the reservoir, interfacial tension is reduced when CO_2 is mixed with oil resulting in increased relative permeability and mobility of the oil. The more mobile the oil is, the easier it is to produce and a higher oil recovery can be achieved (Ghoodjani and Bolouri, 2011).

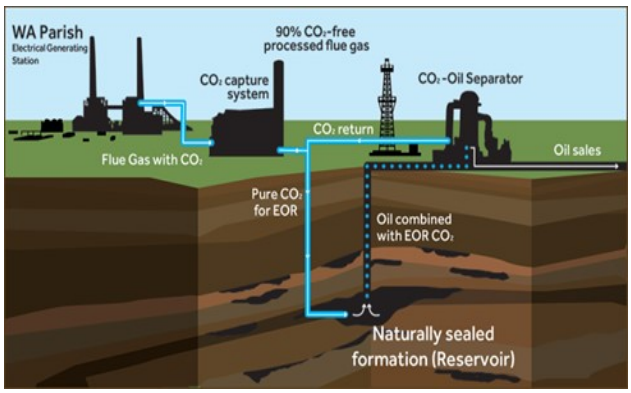


Figure 1. Working principles of CO_2-EOR (Advanced Resources International and Melzer Consulting, 2010).

When CO_2 dissolves in oil, the viscosity of the oil decreases significantly. The reduction of oil viscosity is highly dependent on the initial viscosity of the oil. Less viscous oil will be less affected by the CO_2, while for more viscous oils, the effect of viscosity reduction is more pronounced. The reduction in oil viscosity will cause an increase in the oil relative permeability. This will reduce the residual oil saturation in the reservoir and improve the oil recovery (Ghoodjani and Bolouri, 2011).

CO_2 interacts with the oil in the reservoir and dissolves in the oil at certain reservoir conditions. The dissolution of CO_2 in oil causes the oil to swell. Reservoir characteristics, as pressure and temperature as well as oil composition, determine the strength of the oil swelling effect (Ghoodjani and Bolouri, 2011). Swelling plays an important role in achieving better oil recovery. Variation in the swelling factor influences on the residual oil saturation, which is inversely proportional to the swelling factor. Residual oil saturation, in turn, affects the relative permeability, which plays a crucial role in oil recovery (Ghoodjani and Bolouri, 2011).

Swollen oil droplets force fluids to move out of the pores and oil that initially was unable to move out of the pores under certain pressure conditions will now be forced to move towards the production well. Hence, oil swelling causes drainage effect that decreases the residual oil saturation (Ghoodjani and Bolouri, 2011).

2 Olga and Rocx

OLGA is a one-dimensional transient dynamic multiphase simulator used to simulate flow in pipelines and connected equipment. OLGA consists of several modules depicting transient flow in a multiphase pipeline, pipeline networks and processing equipment. Since Olga is a spatially 1-dimensional simulator, only one set of equations is used for the calculation of the well properties in the length direction. That is, the properties of the fluid are independent of the radius of the well and changes therefore only with length and time (Thu, 2013).

Rocx is a three-dimensional near-well model coupled to the OLGA simulator to perform integrated wellbore-reservoir transient simulations. Rocx can simulate three-phase flow in porous media. It has two OLGA PVT options available, among which black-oil tracking is used in this project. Rocx simulations can be run without the coupling to OLGA. However, by using Rocx in combination with OLGA, more accurate predictions of well start-up and shut-down, observation of flow instabilities, cross flow between different layers, water coning and gas dynamic can be obtained (Schlumberger, 2007). The OLGA simulator is governed by conservation of mass equations for gas, liquid and liquid droplets, conservation of momentum equations for the liquid phase and the liquid droplets at the walls, and conservation of energy mixture equation with phases having the same temperature (Schlumberger, 2007).

2.1 Rocx

Schlumberger (2007) describes the mathematical models used in Rocx in detail. The models for relative permeability developed by Corey and Stone II are presented below.

In this study, the Corey model is used to define the relative permeability curves for water (Li and Horne, 2006; Rocx Online Help). This model is a combination of the Burdine approach for calculation of the relative permeability of the wetting and non-wetting phases and the capillary pressure model that was defined by Corey. The Corey model is also called the Brooks and Corey model depending on the value of the pore size distribution index. If the pore size distribution index is less than 2, the model is called the Corey model and if it is greater than 2, it is called the Brooks and Corey model (Li and Horne, 2006). The Corey model (Rocx Online Help) for predicting the relative permeability of water is given by:

$$k_{rw} = k_{rowc}\left(\frac{S_w - S_{wc}}{1 - S_{or} - S_{wc}}\right)^{n_w} \quad (1)$$

where k_{rw} is the relative permeability of water, k_{rowc} is the relative permeability of water at the maximum water saturation, S_w is the water saturation, S_{wc} is the irreducible water saturation, S_{or} is the residual oil saturation and n_w is the Corey fitting parameter for water.

The Stone II model is used in to calculate the relative permeability of oil. This model is widely used for predicting relative permeability in water-wetted systems with high saturations of oil. The Stone II model estimates the relative permeability of oil in an oil-water system based on the following equation (Rocx Online Help):

$$k_{row} = k_{rowc}\left(\frac{S_w + S_{or} - 1}{S_{wc} + S_{or} - 1}\right)^{n_{ow}} \quad (2)$$

where k_{row} is the relative oil permeability for the water-oil system, k_{rowc} is the endpoint relative permeability for oil in water at irreducible water saturation and n_{ow} is a fitting parameter for oil.

3 Simulation Details

This section describes the simulation method and procedures. A near-well reservoir is constructed in Rocx, imported to OLGA and simulated. The results are presented using OLGA and Tecplot.

3.1 Geometry

The near-well reservoir has a length and width of 60 meter and height of 20 meters. The horizontal base pipe is located in the middle of the x-y plan and 15 meter

above the bottom of the reservoir. Figure 2 presents a schematic overview of the simulated reservoir.

The water drive pressure from the bottom of the reservoir is 320 bar and the pressure in the base pipe is varied from 300 to 317 bar. Differential pressures of 3, 5, 10, 15 and 20 bars were used as the driving force for oil production in the simulation. The pressure at the outer boundary of the reservoir was constant at 320 bars, since the reservoir is considered to be infinitely large. The grid was set to (n_x, n_y, n_z) = (3, 31, 20).

3.2 Reservoir Conditions

Reservoir characteristics were chosen according to the Grane oil field in the North Sea. Grane was selected for this research since this is the first field that started to produce heavy crude oil in Norway (Fath and Pouranfard, 2014). The reservoir is characterized as homogeneous reservoir without gas-cap and with high porosity (0.33) and permeability (up to 10 D). The oil viscosity at reservoir conditions reaches 12 cP with 19°API gravity.

According to Fath and Pouranfard (2014), MMP for carbon dioxide and oil (20° API) is 320 bar (at 121 °C). The reservoir pressure in the Grane field is 176 bar. However, in order to simulate the miscible CO_2-EOR method, the reservoir pressure was set to 320 bar and the temperature to 121 °C. The rock is defined as sandstone, and the thermal properties was chosen from data given by Eppelbaum *et al* (2014). The reservoir characteristics of the Grane field and the reservoir characteristics used in this study are listed in Table 1.

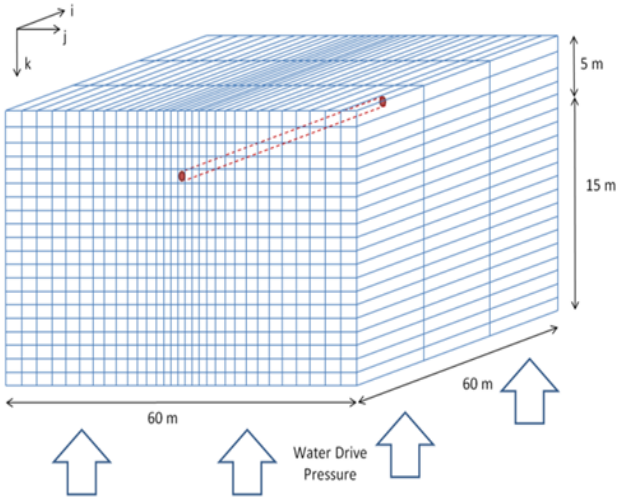


Figure 2. Schematic overview of the near-well reservoir.

Table 1. Reservoir characteristics of Grane field and the simulated reservoir.

Parameter	Grane field reservoir	The simulated reservoir
Oil viscosity	10-12 cp @ 76 °C and 176 bar	12 cp @ 76 °C and 176 bar
Oil specific gravity	0.876 (19° API)	0.876 (19° API)
Porosity	0.33	0.33
Permeability	Up to 10 D	x and y-directions: 7 D z-direction (gravity): 0.7 D
Area	25.5 km^2	1860 m^2 (60X31)
Thickness	31 m	20 m
Gas Oil Ratio	14-18 Sm^3/Sm^3	15 Sm^3/Sm^3
Reservoir pressure	176 bar	320 bar
Reservoir temperature	76 °C	121 °C
Rock compressibility	Not found	0.00001 1/bar
Rock heat conductivity	Not found	1.7 W/mK
Rock heat capacity	Not found	737 J/kgK
Rock density	Not found	2198 kg/m^3
Initial oil saturation	Not found	1
Irreducible water saturation	Not found	0.18
Residual oil saturation	Not found	0.3 (before CO_2 breakthrough)

3.3 CO_2 Injection Simulation

In OLGA, a geometry comprising of two pipelines (one injection and one production flow path), three closed nodes, one pressure node, three pressure sources and six near-well source was used to show the injection of CO_2 and enhanced oil production. This was designed using the basic case function in the OLGA model browser with a pipeline diameter of 0.12 m as shown in Figure 3.

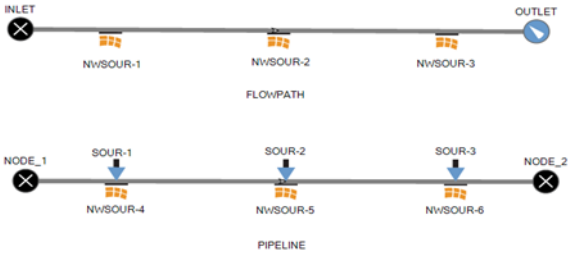


Figure 3. CO_2 injection design in OLGA.

The blackoil model was selected in Rocx and it was assumed that the reservoir was initially saturated with oil. Differential pressure was utilized as a driving force while trying to inject CO2 into the reservoir. This was implemented in OLGA with use of pressure sources and the initial injection well pressure was set to 325 bars, the pressure of the CO2 source was set to 340 bar and the reservoir pressure was set to 320 bar. This condition gave a pressure difference of 15 bar that was considered adequate for injection.

The pressure drive in the reservoir is along the vertical direction, with water coming into the reservoir from the bottom and forcing the oil towards the production well. The injection well was placed close to the bottom while the production well was placed at the upper region of the reservoir as shown in Figure 3. These precautions were taken to ensure effective injection and higher oil recovery.

The use of zones was considered able to inject CO_2 into the near well reservoir. To achieve this, zones (perforations) added along the wellbore to automatically generate inflow in all control volumes in the reservoir. The inflow was expected to be calculated between the boundary positions, as opposed to the well option, which is modeled as a point source. The reservoir pressure and temperature were assumed constant during the whole simulation period and were set to 320 bar and 121 °C respectively. The pressure in the injection well varied between 330 to 400 bar. With the differential injection pressure, it was expected that CO_2 would be injected into the Rocx near well reservoir. However, when using the black oil model, which was the only option for this project, injection of CO_2 directly to the reservoir, was not possible. The further simulations were therefore performed without the injection well, and the effect of CO_2 injection was simulated by changing the relative permeability curves.

4 Results

4.1 Optimum Differential Pressure

Simulations were performed to find the optimum differential pressure for oil production in the actual field. The differential pressure, P_{res} - P_{well}, is the driving force in the oil production, and is mentioned as drawdown. Increasing the differential pressure, increases the oil production rates, but it also increase the risk of early water or gas breakthrough. Figure 4 shows the accumulated oil production over a period of 120 days, when the drawdown was varied from 3 to 20 bar. Figure 4 shows that oil production increases with CO_2 injection. The gradient of the oil production curves indicates that the production rates are highest during the first period of production. After a certain time, the curves are becoming more flat, which shows that the oil production rates are decreasing. This occurs at different time for the different cases, and presents the time of water breakthrough. The accumulated oil production at drawdown 20 bar is about 12000 m^3 after 120 days, whereas the production using 15, 10 and 5 bar drawdown is approximately 11500, 10500 and 8200 m^3 respectively. A drawdown of only 3 bar, gives oil production of 5000 m^3 after 120 days. However, using 3 bar differential pressure, there was no additional oil production after 90 days. This indicates that the drawdown has to exceed 3 bar to get an acceptable oil recovery.

Oil recovery is greatly affected by water breakthrough. Figure 5 presents the water cut (WC) for the different drawdowns. From this figure, the optimal differential pressure that would delay water breakthrough and contribute to higher oil recovery would be determined. When using high drawdown, early water breakthrough occurs. The time of water breakthrough changes from 14 days to 84 days when the drawdown is changed from 20 bar to 3 bar. Due to the high mobility of water compared to oil, the water production increases dramatically after water breakthrough.

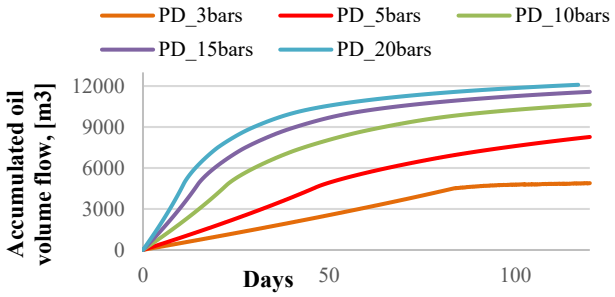


Figure 4. Accumulated oil volume flow for differential pressures of 120 days.

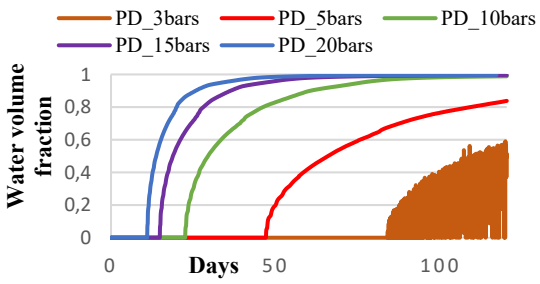


Figure 5. Water cut as a function of drawdown and time.

The Figure 5 shows that when using drawdown of 15 or 20 bars, the water cut reach a value close to unity after very short time. The cases with lower drawdown give a more moderate increase in the WC. However, it can clearly be seen that with 3 bar drawdown, big fluctuations occur. This is mainly due to numerical problems when the drawdown becomes too low. When choosing the optimum drawdown, both the breakthrough time, the water cut and the oil production rate have to be considered.

The oil production rates are presented in Figure 6. It can be seen that the production rates increase until water breakthrough occurs, and then decrease significantly. The gradient for the oil production rate is steepest for the high drawdown cases. The highest oil volume flow rates are reached at the breakthrough time, and were recorded as 360 m^3/day for 20 bar drawdown and as 40 m^3/day for 3 bar drawdown. The peaks of the oil flow rates for 15, 10 and 5 bar differential pressure were 250 m^3/day, 172 m^3/day and 78 m^3/day respectively. At 3 bar differential pressure, the flow became very unstable at the water breakthrough time, which indicates insufficient differential pressure.

Based on the simulation results with different drawdown, 10 bar was considered the optimum drawdown. The reason for choosing 10 bar drawdown, is that the increase in WC with time is much lower than for the 15 and 20 bar cases, which means that oil can be produced at lower separation costs for a longer period of time. In addition, the accumulated oil production after 120 days is only 1500 m^3 lower than for 20 bar, and the well is still producing oil at that time.

Drawdown of 5 bar gives low production rate and 3 bar drawdown gives unstable flow and no oil production after water breakthrough.

Based on this, 10 bar drawdown is used in the further simulations for studying the effect of CO_2 injection on increased oil recovery.

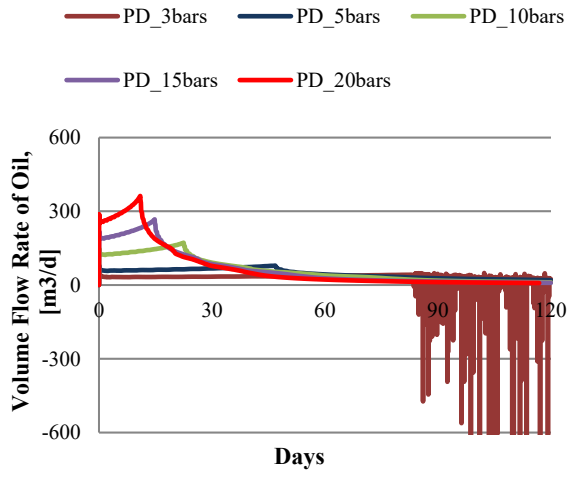


Figure 6. Volumetric flowrate of oil at differential pressures of 120 days.

4.2 CO_2-EOR

As discussed before, injection of CO_2 to the reservoir, changes the relative permeability curves. When CO_2 is injected to the reservoir, it increases the relative permeability of oil and decreases the residual oil saturation. This phenomenon was used in the simulations in order to study the effect of CO_2.

Relative permeability curves were developed in Rocx by using the Corey and STONE II models for water and oil respectively. Model fitting parameters that define the shape of the relative permeability curves were chosen assuming water-wetted rock. The fitting parameter for the water relative permeability curve can be set to 3, and for oil relative permeability curve to 2.5 (Best, 2002). Irreducible water saturation is set to 0.18. In order to simulate the effect of CO_2 injection, the residual oil saturation was decreased from 0.3 to 0.15 with a constant step of 0.05. Residual oil saturation is assumed to be 0.3 without CO_2 injection and 0.15 for effective CO_2-EOR. The relative permeability curves that were calculated are presented in Figure 7. The curves were included in Rocx and used in the further simulations to study the effect of CO_2-EOR. With CO_2 injection, oil relative permeability is increased while the residual oil saturation is reduced. This would enhance oil mobility and improve recovery.

4.3 Accumulated oil volume flow

According to the data presented in Figure 8, the accumulated oil volume flow increases with decreasing residual oil saturation. The accumulated oil production increased from about 10700 m^3 to 12000 m^3 when the residual oil saturation decreased from 0.3 to 0.15. This corresponds to an increased oil production of 12%. The differences in the accumulated oil production with varied residual oil saturation are observed after about 25 days.

This can be explained by the change in relative permeability because of CO_2 injection. At oil saturation between 0.82 and 0.75, the deviation between the relative oil saturation curves are very small, which results in insignificant differences in the oil production. The differences in oil production are getting more pronounced when the oil saturation in the reservoir decreases.

4.4 Accumulated water volume flow

As seen in Figure 9, water breakthrough to the well starts after about 24 days. The accumulated volume of water increases with increasing residual oil saturation. The water production decreased from about 90000 m^3 to 70000 m^3 when the residual oil saturation was reduced from 0.3 to 0.15. This corresponds to 22% water reduction during the 120 days of production. Hence, the higher the residual oil saturation, the higher is the total water production. This can be explained based on the relative permeability curves that are shown in Figure 7.

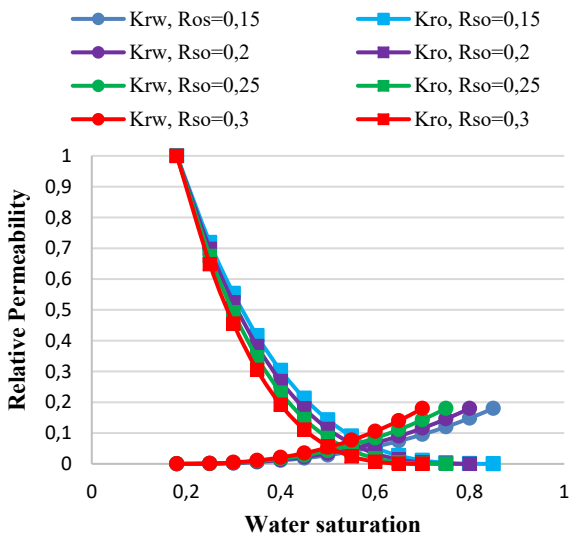


Figure 7. Tested relative permeability curves.

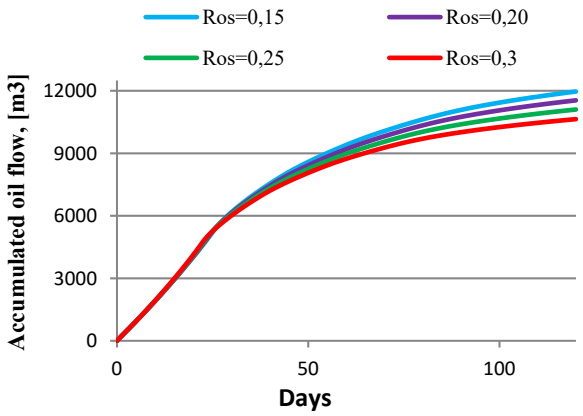


Figure 8. Accumulated oil volume flow.

When the residual oil saturation is changed the water relative permeability curve also changes. These changes are more pronounced at higher water saturations. Therefore, a decrease in residual oil saturation contributes not only to an increase in oil production, but also to a decrease in the water production.

4.5 Oil saturation in the reservoir

This study includes near well simulations, meaning that only a limited part of the reservoir is considered. Initially the reservoir contained 100% oil. After 120 days of production, the oil saturation has decreased significantly in the near well area. Figure 10 represents the reservoir saturation after 120 days when the CO_2 is injected to the reservoir and the residual saturation of oil is 0.15. The plot shows how the water from below is coning towards the well, replacing the oil. The saturation of oil varies with location in the reservoir, and the highest saturation is found above the well location.

The minimum, average and maximum oil saturation in the reservoir after 120 days are estimated and the results are presented in Figure 11. The average oil saturation in the reservoir is decreasing from 48% to 41% when the residual saturation changed from 0.3 to 0.15. The oil recovery is defined as the ratio of produced oil to the original oil in place (OOIP). In this stud, it was assumed that the initial oil saturation was 100%. The oil recovery after 120 days of production is 52%, 54%, 57% and 59% for the cases with residual oil saturation of 0.3, 0.25, 0.20 and 0.15 respectively.

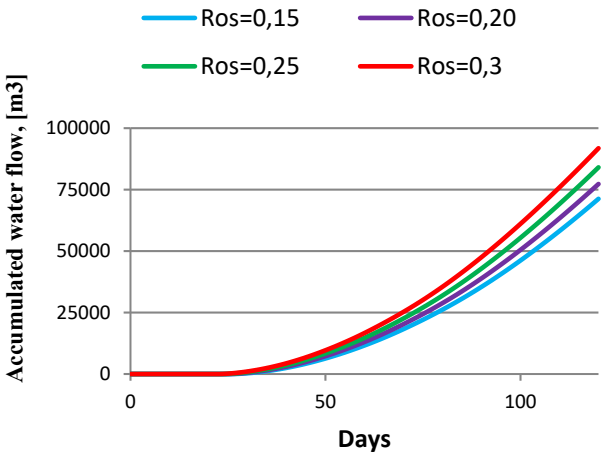


Figure 9. Accumulated water volume flow.

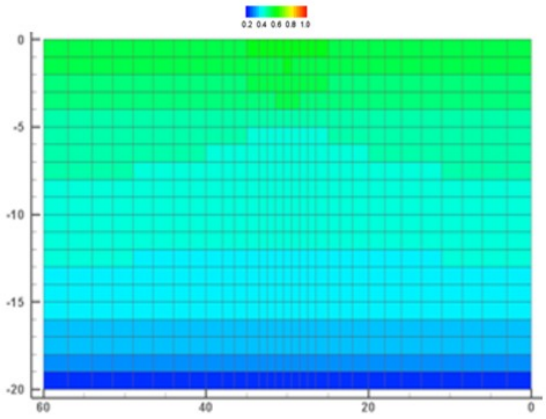


Figure 10. Reservoir YZ-profile (width-height) after 120 days of production using S_{or}=0.15.

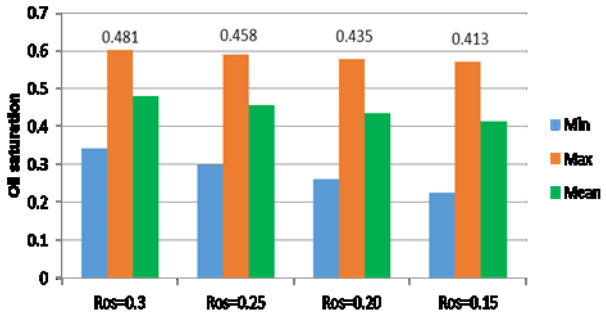


Figure 11. Oil saturation distribution in reservoir after 120 days.

5 Conclusions

Simulation of oil production is performed using the near-well software Rocx in combination with OLGA. Different models were developed to simulate CO_2 injection into the reservoir. However, the authors did not succeed in simulating CO_2 injection when the black-oil model was used, and the effect of CO_2 was studied by changing the relative permeability curves. Preliminary simulations were performed to determine the optimal drawdown for the oil production. The optimal drawdown was selected as 10 bar, because 10 bar gives stable production, relatively high oil production and acceptable water breakthrough time. A series of simulations were performed to evaluate the effect of CO_2 injection on oil recovery by adjusting the relative permeability curves using the Corey and STONE II correlations. The simulation results show that when the residual oil saturation is decreased from 0.3 to 0.15 due to CO_2 injection, the oil production increases with 12%, the water production decreases with 22%, the water cut decreases from 88% to 86% and the oil recovery factor increases from 0.52 to 0.59. This confirms that CO_2 is well suited for enhanced oil recovery.

References

Advanced Resources International and Melzer Consulting. *Optimization of CO2 Storage in CO2 Enhanced Oil Recovery*. Projects, prepared for UK Department of Energy and Climate Change, 2010. Available via: http://neori.org/resources-on-co2-eor/how-co2-eor-works/ [accessed September 2016].

A. H. Fath and A. R. Pouranfard. Evaluation of miscible and immiscible CO2 injection in one of the Iranian oil fields. *Egyptian Journal of Petroleum* 23(3):255-270, 2014.

E. Ghoodjani and S. Bolouri. Experimental study and calculation of CO2-oil relative permeability. *Petroleum and Coal* 53(2):123-131, 2011.

E. S. Thu. *Modeling of Transient CO2 Flow in Pipelines and Wells*. Master`s thesis, Institutt for energy og prosessteknikk, 2013.

K. D. Best. *Development of an integrated model for compaction/water driven reservoirs and its application to the J1 and J2 sands at Bullwinkle, Green Canyon Block 65, Deepwater Gulf of Mexico*. Doctoral Dissertation, Pennsylvania State University, 2002.

K. Li, and R. N. Horne. Comparison of methods to calculate relative permeability from capillary pressure in consolidated water-wet porous media. *Water resources research* 42(6):1-9, 2006.

L. Eppelbaum, I. Kutasov. and A. Pilchin. Thermal properties of rocks and density of fluids. *In Applied geothermics*:99-149, 2014.

L. S. Melzer. *Carbon dioxide enhanced oil recovery (CO2 EOR): Factors involved in adding carbon capture, utilization and storage (CCUS) to enhanced oil recovery.* Center for Climate and Energy Solutions, 2012. Available via: http://neori.org/Melzer_CO2EOR_CCUS_Feb2012.pdf [accessed September 2016].

L. Zhang, B. Ren, H. Huang, Y. Li, S. Ren, G. Chen, and H. Zhang. CO2 EOR and storage in Jilin oilfield China: monitoring program and preliminary results. *Journal of Petroleum Science and Engineering* 125:1-12, 2015.

Rocx Help. Reservoir characteristics. [Online].

Schlumberger. Rocx Reservoir Simulator, 1.2.5.0 edition, 2007.

T. A. Jelmert, N. Chang, L. Høier, S. Pwaga, C. Iluore, Ø. Hundseth, and M. U. Idrees. *Comparative Study of Different EOR Methods*. Norwegian University of Science and Technology, Trondheim, Norway, 2010.

Simulation of Heavy Oil Production using Inflow Control Devices

A Comparison between the Nozzle Inflow Control Device and Autonomous Inflow Control Device

Emmanuel Okoye , Britt M. E. Moldestad
Department of Process, Energy and Environmental Technology, University of Southeast Norway, Porsgrunn, Norway.
Britt.Moldestad@usn.no, Chuksokoye2@gmail.com

Abstract

Production of heavy oil requires the application of new technologies in order to handle the challenges associated with the production. The main challenges are early water breakthrough, resulting in high water cut and low oil recovery. Especially in heterogeneous reservoirs, early water breakthrough and high water cut lead to low productivity and high separation costs. Different types of inflow control devices (ICDs) have proven to be effective in delaying water breakthrough and the newer technology has also the ability to choke for water after breakthrough. The near well simulation tool, NETool, was used to simulate oil production from homogeneous and heterogeneous heavy oil reservoirs after water breakthrough has occurred. The oil and water production, using nozzle ICD and autonomous ICD (RCP) completion, has been simulated and compared. ICD is producing more oil than RCP, but it is also producing significantly more water. The well with ICD completion gave about 4 to 5 times higher water cut than the well with RCP completion. Estimates indicate that by utilizing the newest technology, autonomous inflow control valve (AICV), the water cut can be reduced significantly without reducing the oil production.

Keywords: ICD, RCP, AICV, water cut, water breakthrough, heavy oil

1 Introduction

Heavy oil is a type of unconventional oil with a viscosity above 100cP and API gravity less than 22.3°. Heavy oil has in the past years been used as a source of refinery feedstock due to its lower quality compared to conventional oil (Meyer *et al*, 2007). Due to a high population growth and a massive decline in conventional oil reserves in recent times, there has been a need of developing advanced technologies to improve heavy oil recovery. Inflow control devices as an advanced technology have the ability to reduce the water-cut and improve oil recovery. Early water breakthrough poses a real challenge in the recovery of heavy oil in reservoirs with water drive. An estimate gives that about 70% of heavy oil is left behind in the reservoir after the production is shut down since it is no longer economical to produce oil (Aakre *et al*, 2014).

For both homogenous and heterogeneous reservoirs, high differential pressures are needed to aid production of heavy oil. This leads to an early water breakthrough to the well. The inflow control devices, on the other hand, have the potential of delaying water breakthrough, and as a consequence the production can run with a higher drawdown to increase the production rate (Aakre *et al*, 2014).

In this study, a comparison between two different types of inflow control devices namely the Nozzle inflow control device (ICD) and the Autonomous inflow control device (RCP) is considered using the NETool reservoir simulator. Simulation of oil flow rates, water flow rates and water-cut (WC) at various drawdown pressures are carried out. Result analysis, discussion and conclusions are given.

2 Inflow control devices

2.1 Nozzle inflow control device

The nozzle ICDs are often used in long horizontal wells and to balance the production rates between different zones of the well. ICD has the ability of delaying early water breakthrough, thus reducing the average water cut in the well. Moreover, the ICD is passive – meaning it neither chokes nor closes after water breakthrough has occurred (Aakre *et al*, 2014). The principle behind the nozzle ICD is based on the following equations (Halliburton, 2014):

$$\Delta P = \frac{\rho v^3}{2C^2} = \frac{\rho Q^2}{2A_{valve}^2 C^2} = \frac{8\rho Q^2}{\pi^2 D_{valve}^4 C^2} \quad (1)$$

$$C = \frac{C_D}{\sqrt{(1-\beta^4)}} = \frac{1}{\sqrt{K}} \quad (2)$$

$$\beta = \frac{D_2}{D_1} \quad (3)$$

where ΔP is pressure drop across orifice, ρ is average fluid density, v is fluid velocity through orifice, Q is fluid flow rate through orifice, A is area of orifice, D is

diameter of orifice, C is flow coefficient, C_D is discharge coefficient, and K is pressure drop coefficient.

2.2 Autonomous inflow control device

The Autonomous Inflow Control Device (AICD) used in this study, is known as Rate Controlled Production (RCP) and is developed by Statoil. The view of the RCP is to delay water breakthrough and autonomous chocking of water after water breakthrough. Autonomous, means that the inflow control device is self-regulating and it is not controlled from the surface. This autonomous behavior enables the RCP to produce more heavy oil from the long horizontal wells (Mathiesen *et al*, 2011) by chocking the zones that are producing water, and at the same time produce oil from the other zones. This implies that high drawdown can be used, and the oil recovery can be increased significantly. The performance of the RCP is based on Bernoulli's equation:

$$P_1 + \frac{1}{2}\rho V_1^2 = P_2 + \frac{1}{2}\rho V_2^2 + \Delta P_{Friction\ loss} \quad (4)$$

Where P_1 and P_2 are static pressures, the velocity terms represent the dynamic pressures, $\Delta P_{Friction\ loss}$ is pressure loss due to friction and ρ is density of the fluid.

The RCP is characterized by being very little sensitive to changes in differential pressure, and gives a more uniform flow rate over a range of drawdowns compared to the ICD. The following equations describe the functionality of the RCP:

$$\delta P = f(\rho,\mu) \,.\, a_{AICD} \,.\, q^x \quad (5)$$

$$f(\rho,\mu) = \left(\frac{\rho_{mix}^2}{\rho_{cal}}\right) \cdot \left(\frac{\mu_{cal}}{\mu_{mix}}\right)^y \quad (6)$$

$$\rho_{mix} = \alpha_{oil}\,\rho_{oil} + \alpha_{water}\rho_{water} + \alpha_{gas}\rho_{gas} \quad (7)$$

$$\mu_{mix} = \alpha_{oil}\mu_{oil} + \alpha_{water}\mu_{water} + \alpha_{oil}\mu_{oil} \quad (8)$$

Where δP is pressure drop through RCP, q is the flow rate, x and y are user input constants, a_{AICD} is the valve strength parameter, α is the volume fraction of the actual phase, ρ_{cal} and μ_{cal} are calibration density and viscosity.

2.3 Autonomous inflow control valve

Autonomous Inflow Control Valve (AICV) is developed by Inflow Control AS. Unlike other inflow control devices that delay and/or choke the water production, the AICV closes completely when water breakthrough occurs and re-opens again when the oil is well saturated around the valve. The AICV technology can be used in long horizontal wells for various type of oil production, ranging from light oil to bitumen production using SAGD. The device is said to be self-regulating and gives very low restriction for oil flow (Aakre *et al*, 2014). The working principle of the AICV is based on the difference in pressure drop through a laminar and a turbulent flow restrictor. The pressure drop through the laminar flow restrictor is proportional to the viscosity and the velocity and is expressed by the equation for pressure drop through a pipe segment (Aakre *et al*, 2014). The pressure drop through the turbulent flow restrictor is proportional to the density and the velocity squared. The pressure drops through the laminar and turbulent flow restrictors are expressed by eq. (9) and (10) respectively (Aakre *et al*, 2014):

$$\Delta P = f \cdot \frac{L\cdot\rho\cdot v^2}{2D} = \frac{64}{Re} \cdot \frac{L\cdot\rho\cdot v^2}{2D} = \frac{32\cdot\mu\cdot\rho\cdot v\cdot L}{D^2} \quad (9)$$

$$\Delta P = k \cdot \frac{1}{2} \cdot \rho \cdot v^2 \quad (10)$$

Where ΔP is pressure drop, f is friction coefficient, ρ is the fluid density, μ is fluid viscosity, L is length of laminar flow element, D is diameter of laminar flow element, Re is Reynolds number, k is geometrical constant and v is fluid velocity.

Since AICV can close for water and gas, it has a very high potential for increased oil recovery. However, the AICV is a new development and is still not included as an option in NETool. The simulations are therefore focused on the nozzle ICD and the RCP, and the potential of the AICV will be discussed based on the results from these two types of inflow controls.

3 NETool

NETool is a steady state near well simulation tool, and can be used for analysis of the effect of different completion components, near-wellbore effect on productivity of the well, modeling of completion components in the production and injection interval, design of inflow control devices to delay water and gas breakthrough, etc. NETool can also be linked with other software like design software and reservoir simulators such as Eclipse and Nexus. The black oil model is included in NETool, and is used in all the simulations.

3.1 Input parameters

In the simulations, two different cases were considered; one case for homogenous reservoir and one case for heterogeneous reservoir. The same reservoir properties were utilized for both reservoir types except for the reservoir permeability as shown in Table 1.

Table 1. Reservoir input parameters

Reservoir Parameters and well specifications	Case 1 Homogeneous Reservoir	Case 2 Heterogeneous Reservoir
Well length	800m	800m
Reservoir thickness	100m	100m
Reservoir Width	2500m	2500m
Reservoir Pressure	300bar	300bar
Porosity	0.2	0.2
Permeability	2000 md	2000md & 10000md
Oil viscosity	43.7cP	43.7cP
API gravity	18	18
Reservoir Temperature	24.87	24.87
Water gravity	1000 kg/Sm^3	1000 kg/Sm^3
Dissolved gas/oil ratio	80 Sm^3/Sm^3	80 Sm^3/Sm^3

A simple well completion was used, as more emphasis was placed on contrasting between the two inflow control devices, ICD and RCP. The well completion data is shown in Table 2.

Table 2. Well completion

Well completion parameters	Case 1 Homogeneous Reservoir	Case 2 Heterogeneous Reservoir
No of well segments	32	32
Length of well segment	25m	25m
No. of Packers	1	4
Total no. of ICDs/RCPs	62	56
No. of ICDs/RCPs producing water	8	8

3.2 Relative permeability

The relative permeability of a given fluid is the ratio of the effective permeability at a particular saturation to the absolute permeability. The relative permeability is used to predict the movement of oil, water and gas in the reservoir. The velocity of fluids flowing in the reservoir are dependent on the relative permeability (Dake, 1978). The reservoir used in the simulations was assumed to be water-wetted. Furthermore, Corey's model and the Stone II model were used in the determination of the relative permeability curves for water and oil. The relative permeability curves for the particular reservoir has to be specified in NETool. The estimated relative permeability curves are presented in Figure 1.

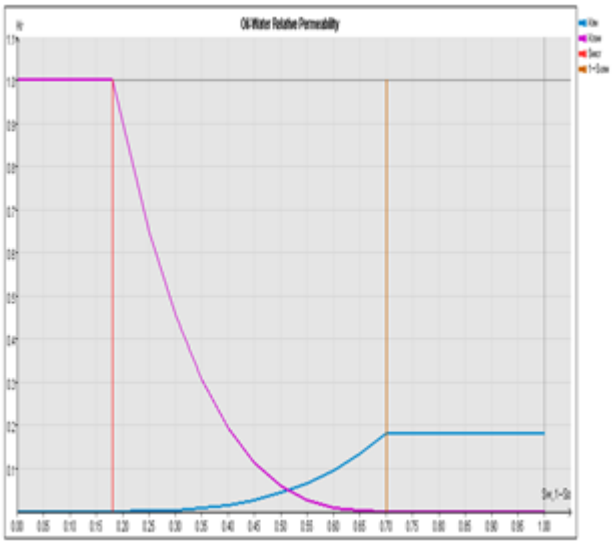


Figure 1. Relative permeability curve for water-wetted reservoir

3.3 Tuning of performance curves

To obtain updated performance curves for ICD and RCP, the constants in the equations were adjusted to fit the production rates as a function of pressure to experimental data. The functionality of the nozzle ICD is well-known, and the default values in NETool gave good fit to experimental data. The discharge coefficient for ICD was set to 0.79. Experimental data (Halvorsen *et al*, 2012; Mathiesen *et al*, 2011) were used to tune the RCP user input parameters x and y. The values of x and y were found to be 3.8 and 1.1 respectively.

3.4 Homogeneous Reservoir

When simulating the homogeneous reservoir, the reservoir parameters in Table 1 were utilized. The permeability of the reservoir was taken to be 2000mD as shown in Figure 2. Some zones in the near-well reservoir were assumed to have 100% oil saturation while other zones had 100% water saturation. At the heel area of the well, the water saturation in the first two zones was set to 100%. A drawing of the well with completion and packer is presented in Figure 3. Packers are used for zonal isolation, in order to prevent water flowing to other zones through the annulus. ICD and RCP completion was used when simulating oil production at different drawdown pressures. The drawdown was set as 5bar, 10bar, 15bar, 20bar, 25bar, 30bar. The diameter of the nozzle ICD was given as 5.0mm while the input strength parameter for the RCP was 1.0e^{-5}. The rest of the completion parameters used in the simulations can be found in Table 2.

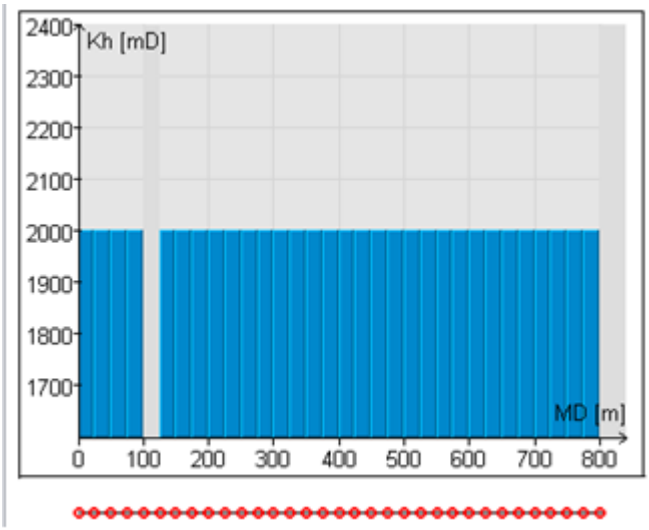


Figure 2. Permeability of homogeneous reservoir

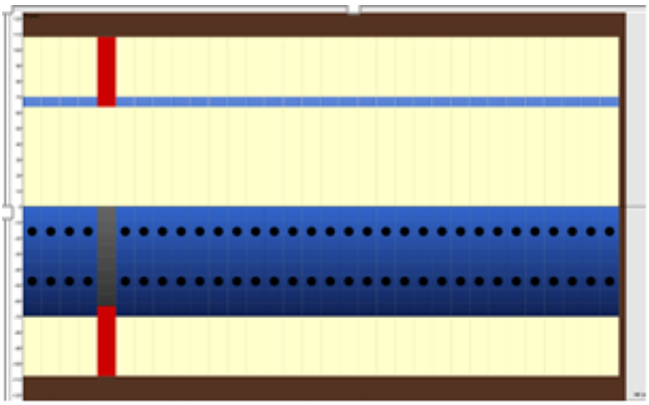


Figure 3. A sketch of the well with inflow controllers (black dots) and packers (red rectangels).

3.5 Heterogeneous Reservoir

In the case of the heterogeneous reservoir simulation, the permeability was set to 10000mD in two near-well zones and 2000mD in the rest of the zones as shown in Figure 4. The zones with high permeability were assumed to have water saturation of 100% while the zones with 2000mD had oil saturation of 100%. Zonal isolation using packers on both sides of the high permeability zones were applied. Figure 5 shows the well with completion and packers. Several drawdown pressures ranging from 5 to 30 bar were used to show the comparative strengths of the ICD and RCP. In addition, the percentage water-cut was also compared for the two types of inflow control devices.

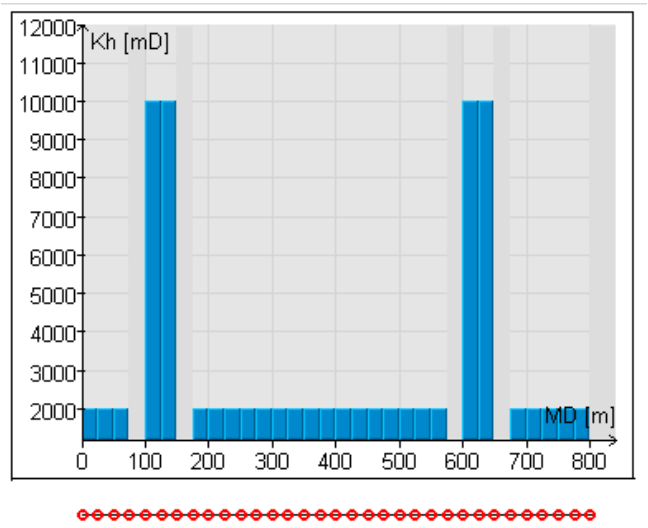


Figure 4. Permeability of heterogeneous reservoir

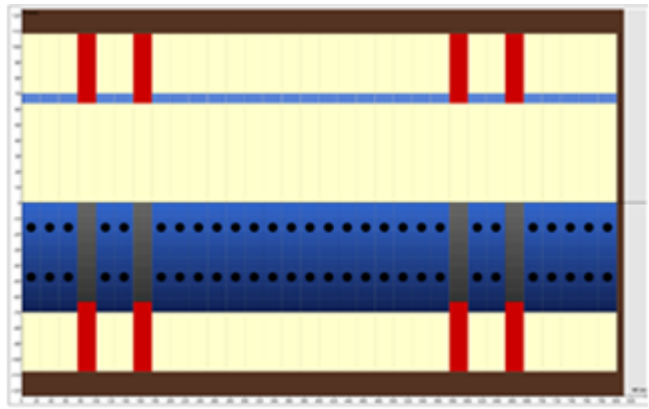


Figure 5. A sketch of the well with inflow controllers (black dots) and packers (red rectangels)

4 Simulation results and discussion

In this section, the results of the two simulation cases with varying drawdown pressures will be analyzed and discussed

4.1 Homogeneous Reservoir

Figures 6 and 7 show the pressures drop through the completion plotted against the oil and water flowrate per RCP and ICD. In Figure 6 the comparison between oil production through ICD and RCP is presented. As can be seen, RCP has a much steeper performance curve than ICD. The ICD has a strength of 2.2 bar. The ICD strength is defined as the pressure drop over the ICD when producing 1 m^3 of oil. By extrapolating the RCP curve, the strength is found to be about 17.5 bar. The strength of the inflow control, influence on the time of water breakthrough. It is also possible to produce with a higher drawdown when the inflow control has a high strength. Since NETool is a steady state simulator, the time of water breakthrough cannot be estimated. The simulations were therefore run assuming that the water breakthrough had already occurred. Figure 7 shows that

the water production through ICD is significantly higher than through RCP. This can also clearly be seen in Figure 8 where the water-cut is plotted as a function of drawdown for RCP and ICD. The water-cut using RCP completion decreases from 14% to 6% when the drawdown is increased from 5 to 30 bar, whereas the water-cut using ICD completion decreases from 58% to 38% when the drawdown is increased from 5 to 30 bar. In Figure 9, the total production rates for oil and water are presented. At drawdown above 10 bar, the ICD are producing more oil, but also significantly more water than the RCP. ICD and RCP are both producing 450 m^3/h of oil at drawdown 10 bar, but at this pressure ICD is producing 10 times more water than RCP. These results indicates that a well with RCP completion can improve the oil production and increase the oil recovery.

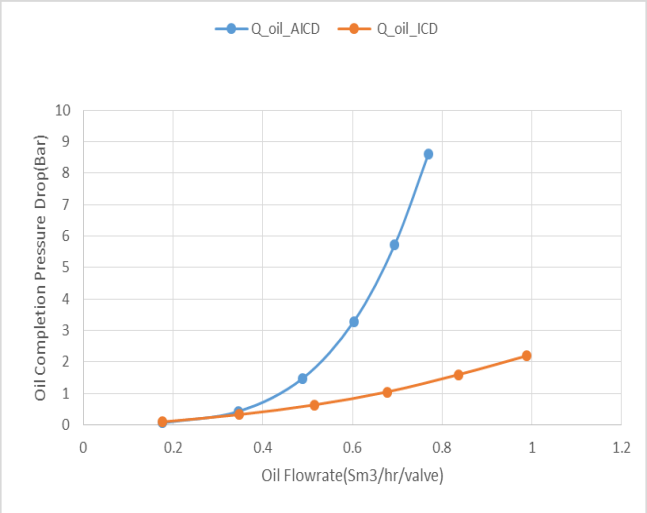


Figure 6. The plot shows the oil completion pressure relative to the oil flowrate per valve for AICD and nozzle ICD

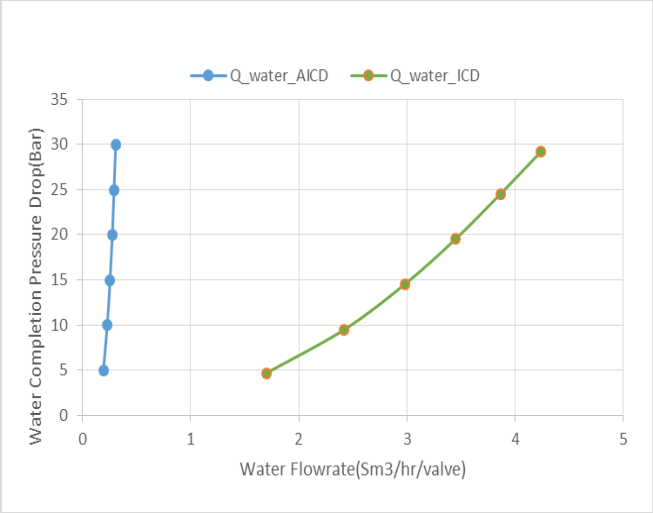


Figure 7. Comparison between the AICD and nozzle ICD showing the plot of water completion pressure versus water flowrate

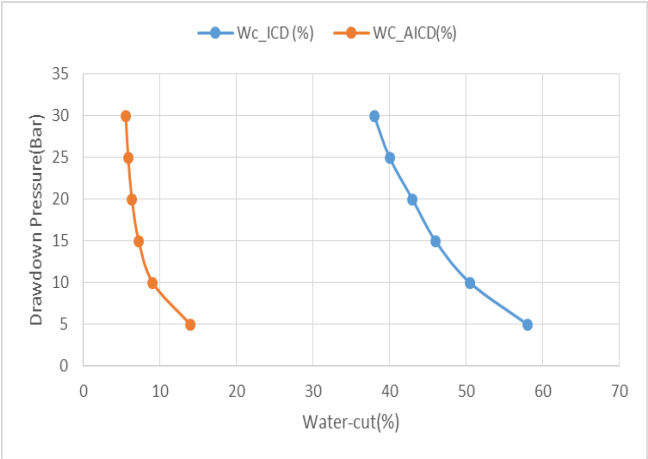


Figure 8. Comparison of water-cut between the AICD and nozzle ICD

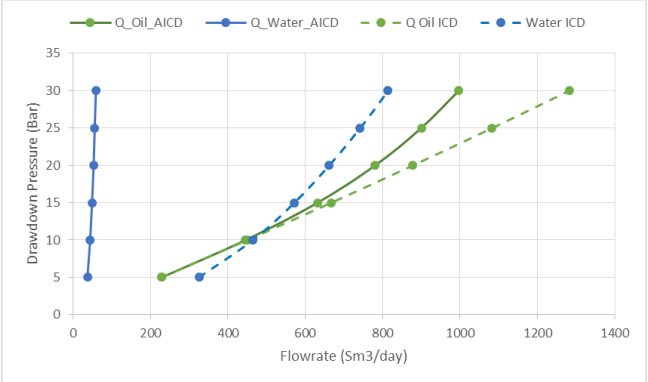


Figure 9. Total production rates as a function of drawdown; homogeneous reservoir.

4.2 Heterogeneous Reservoir

The two figures, Figures 10 and 11, represent the production rate of oil and water for the different inflow control devices with respect to the completion pressure drops. The pressure drop over RCP is proportional to the volume flow rate of oil in the power of about 4 (q^4), whereas the pressure drop over ICD is proportional to the volume flow rate squared (q^2). The strength of the RCP is approximately 10 times the ICD strength. The valve strength of the inflow control devices are calculated by taking the completion pressure corresponding to 1 m^3/h oil production. Figure 11 shows that the production rate of water per ICD is considerably higher than the water flow rate through RCP. At 15 bar differential pressure, the ICD is producing 12 times more water than the RCP.

The water-cut presented in Figure 12 clearly shows that the RCP has a significantly lower water-cut compared to the ICD. Both devices show a decreasing water-cut with increasing drawdown. The water-cut is defined as the ratio of water to total liquid in the well. It is important to have in mind that the water cut shown here is based on water production through 8 of 56

ICDs/RCPs. With time, more and more of the ICDs/RCPs will start to produce water, and the water cut will increase dramatically, especially for the ICD wells. Since the RCP is able to choke for water, RCP completed wells will be able to produce oil for a much longer period before they have to shut down due to high water-cut.

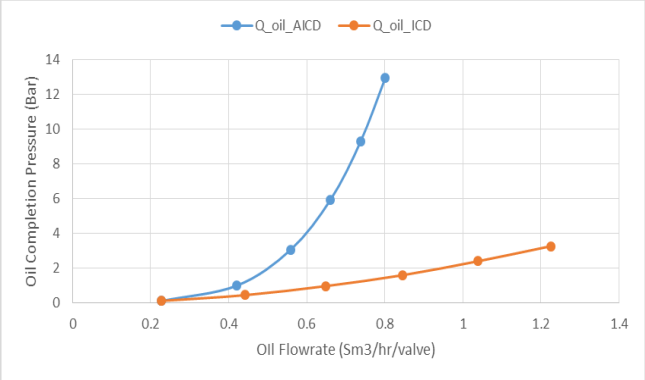


Figure 10. Plot shows the oil completion pressure relative to the oil flow rate per valve for AICD and nozzle ICD.

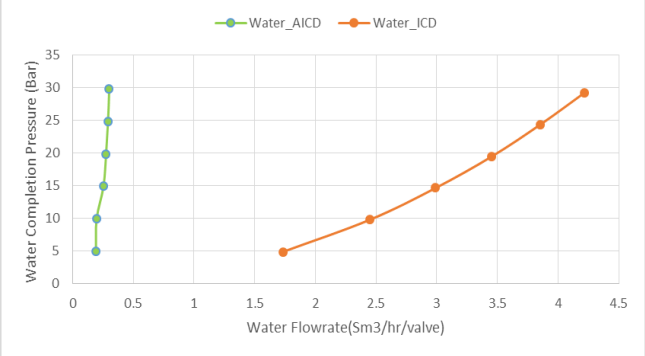


Figure 11. Comparison between the AICD and nozzle ICD showing the plot of water completion pressure versus water flowrate

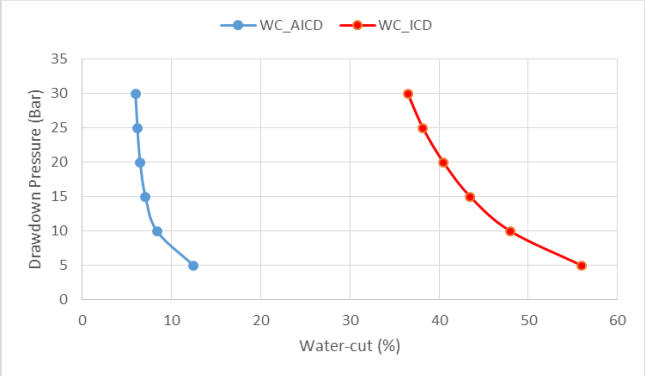


Figure 12. Comparison of water-cut between AICD and nozzle ICD.

Figure 13 show the total production of oil and water from RCP and ICD completed wells at drawdowns ranging from 5 to 30 bar. The flow rate is given in Sm^3/day. At 10 bar drawdown, the production rate of oil is about 500 Sm^3/day for both types of completion. The water production at the same drawdown is about 40 and 470 Sm^3/day for the RCP and ICD completed wells respectively. These results show that wells with RCP completion have a high potential to improve the oil recovery also from heterogeneous reservoirs with high permeable zones where water breakthrough can occur after very short time of production.

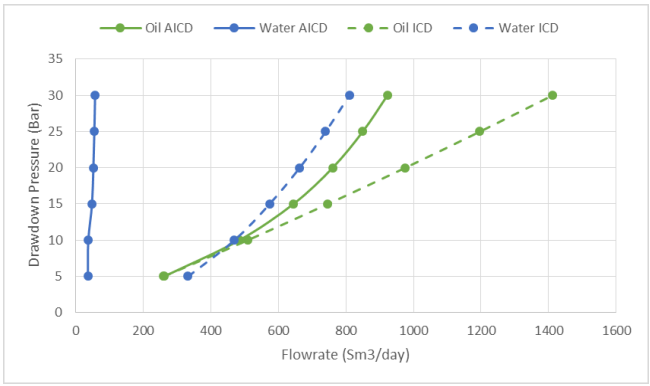


Figure 13. Total production rates as a function of drawdown; heterogeneous reservoir.

4.3 Comments to simulations and results

The simulations have shown that NETool is able to predict oil and water production for different types of well completion. Oil and water production is dependent on the permeability and the relative permeability in the reservoir. It is therefore crucial to have knowledge about the reservoir properties in order to estimate the most appropriate permeability curves for the different types of reservoirs. Production data are needed to tune the relative permeability curves for the particular reservoir. Experimental data are also needed to get a good prediction of the functionality of the RCP and ICD.

AICV completion has not been simulated in this study. However, the functionality of AICV, is that it acts as an ICD when it is surrounded by oil, and closes down to less than 1% of the flow area when it is surrounded by water. This means that AICV is able to produce high amounts of oil and negligible amounts of water after water breakthrough. Since the AICV closes off zones with water, a higher drawdown can be used, and oil can be produced with higher production rates. A rough estimate gives that at 25 bar drawdown, a well with AICV completion in the heterogeneous reservoir has the potential to produce 1200 m^3 oil and about 10 m^3 water per day. Simulations and experimental research confirming the potential of AICV for increased oil recovery are presented by Aakre et al. (Aakre *et al*, 2013; Aakre *et al*, 2013).

5 Conclusions

The near well simulation tool NETool have been used to simulate oil production from homogeneous and

heterogeneous heavy oil reservoirs after water breakthrough has occurred. Inflow controls, RCP and ICD, and packers are used to reduce the water production. RCP has the ability of choking the water rate after breakthrough, whereas ICD only delay the water-breakthrough. NETool is a steady state 1-dimensional simulator, and it cannot predict the time of water breakthrough. However, the strength of the ICD and RCP indicates how much these two devices will restrict the production and thereby delay the time of breakthrough. The oil and water production using ICD and RCP completion have been simulated and compared. ICD is producing more oil than RCP, but it is also producing significantly more water. The well with ICD completion gave about 4 to 5 times higher water cut than the well with RCP completion. Simulations of homogeneous and heterogeneous reservoirs gave about the same results. Estimates indicates that by utilizing AICV completion the water cut can be reduced significantly without reducing the oil production.

References

H. Aakre, B. Halvorsen, B. Werswick, V. Mathiesen. Autonomous Inflow Control Valve for Heavy and Extra-Heavy Oil. *SPE 171141, SPE Heavy and Extra Heavy Oil Conference - Latin America, Medellin, Colombia, 24–26 September, 2014.*

H. Aakre, B. Halvorsen, B. Werswick, V. Mathiesen. Smart well with autonomous inflow control valve technology. *SPE 164348-MS, SPE Middel East Oil and Gas Show and Exhibition, Manama, Barhain, March, 2013.*

H. Aakre, B. M. Halvorsen, B. Werswick, V. Mathiesen. Increased oil recovery of an old well recompleted with Autonomous Inflow Control Valve (AICV). *ADIPEC 2013 Technical Conference, Abi Dhabi, UAE, November10-13, 2013.*

L. P. Dake, Fundamentals of Reservoir Engineering. Amsterdam. Developments in petroleum science 8, Elsevier, 1978.

M. Halvorsen, O. M. Nævdal, G. Elseth. Increased oil production by autonomous inflow control with RCP valves. *SPE 159634, SPE Annual Technical Conference and Exhibition. San Antonio, Texas, USA, October, 2012.*

V. Mathiesen, H. Aakre, B. Werswick, G. Elseth. The Autonomous RCP Valve-New Technology for Inflow Control in Horizontal Wells. *SPE Annual Technical Conference and Exhibition, Aberdeen, UK, 2012.*

R. F. Meyer, E. D. Attanasi P. A. Freeman. Heavy oil and natural bitumen resources in geological basins of the world. Available: http://pubs.usgs.gov/of/2007/1084, 2007.

L. s. services, NETool 5000.0.4.X Technical Manual. *Halliburton, 2014.*

Modeling of Wood Gasification in an Atmospheric CFB Plant

Erik Dahlquist[1], Muhammad Naqvi[1], Eva Thorin[1], Jinyue Yan[1], Konstantinos Kyprianidis[1], Philip Hartwell[2]

[1]School of Sustainable Development of Society and Technology, Mälardalen University (MDH), Sweden,
{erik.dahlquist,raza.naqvi, eva.thorin, jinyue.yan,
konstantinos.kyprianidis}@mdh.se
[2]BioRegional MiniMills (UK) Ltd., United Kingdom

Abstract

The energy situation in both process industries and power plants is changing and it is of interest to investigate new polygeneration solutions combining production of chemicals with the production of power and heat. Examples of such chemicals are methane, hydrogen, and methanol etc. Integration of gasification into chemical recovery systems in the pulp and paper production systems and into the combined heat and power (CHP) systems in power plant applications are among the possible polygeneration systems. It is also interesting to look at the potential to introduce combined cycles with gas turbines and steam turbines as a complement. To perform such analysis, it is important to have relevant input data on what gas composition we can expect from running different type of feed stock. In this paper, we focus on the wood pellets. Experimental results are correlated into partial least squares models to predict major composition of the synthesis gas produced under different operating conditions. The quality prediction models then are combined with physical models using Modelica for investigation of dynamic energy and material balances for large plants. The data can also be used as input to analysis using e.g. ASPEN plus and similar system analysis tools.

Keywords: wood pellets, gasification, CHP, Modelica, methane, hydrogen

1 Introduction and Literature Review

In the following section, short style guidelines are given. In this paper modeling of biomass gasification systems is discussed especially the approach based on energy and mass balances in combination with partial least square (PLS) models developed from experimental results. Dipal and Baruah (2014) made an overview of biomass gasification modelling. Different modeling approaches were categorized based on criteria such as type of gasifier, feedstock, modeling considerations and evaluated parameters. Gómez-Barea and Leckner (2010) structured the modeling work performed with approaches, from artificial neural nets to computational fluid dynamics. The study covered the conversion of fuel particles, char, and the gas and concluded that most of the different approaches fit quite well between models and the experimental results. However, there are research knowledge gaps exist in case of real gasifiers or systems at a large scale. Capata and Veroli (2012) made a mathematical model over an air-blown a circulating fluidized bed (CFB) gasifier with a capacity of 100 kWth range. The study concluded that there were some problems to get reasonable predictions of tar formation. It is interesting to note that we did not create any detectable amounts of tar when we were running our CFB gasifier with a capacity of 100-200 kWth with the wood pellets. This shows that the type of fuel and plant operating conditions affect the gasification results. It becomes difficult to obtain accurate models correlating to the experiments unless the mechanisms are not completely understood. Blasi (2015) has made an overview of the kinetic processes in detail to describe tar formation from a theoretical perspective. Still, it is important to describe what is actually taking place inside the reactors to be able to predict the process.

2 Description of the Pilot Plant

The experimental work has been performed in a pilot plant at Bioregional mini-mills in Manchester. The CFB reactor is heated with a combustor, which is shut off when the operating temperature is achieved. The pressurized air is heated in an electric heater to the desired operating temperature. The gas produced in the CFB is cleaned in two ceramic filters in parallel followed by a scrubber. The gas analysis is performed using an ABB gas chromatograph (GC). The gas is extracted continuously and a new sample is introduced to the GC approximately every five minutes. The plant lay out is shown in Fig. 1 that includes the gasifier, cyclone, gas filtration unit, biomass feeder, hot gas generator and the GC.

In Table 1, we have presented the experimental results from the pilot plant. The feed rate is in ton DS/h.m^2 based on the reactor size. The relative oxidation (Relox) means the amount of air (m^3) needed for the 100 % oxidation of 1 kg of fuel (dried solids).

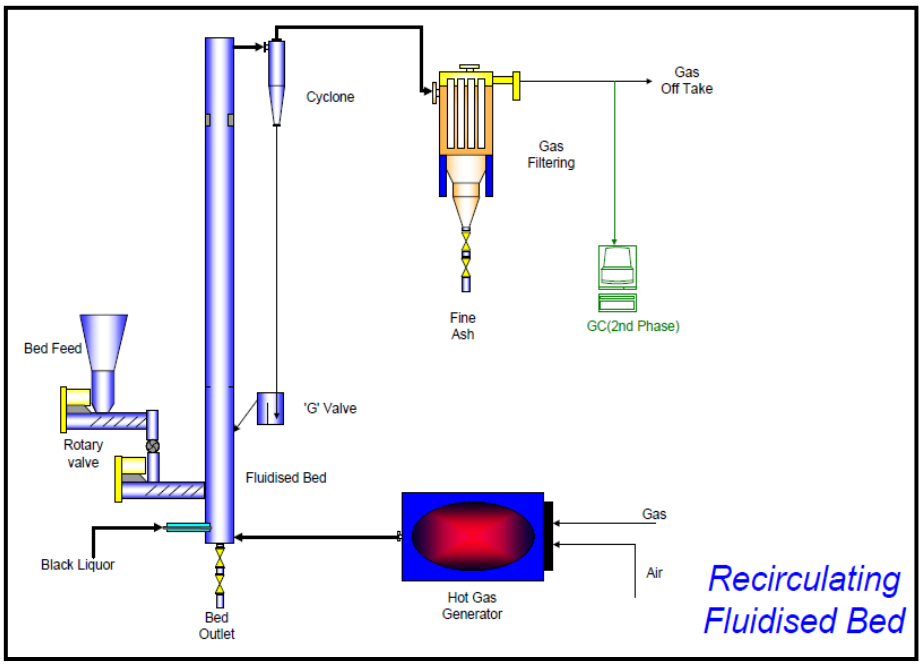


Figure 1. Schematic illustration of the CFB gasifier

Table 1. Results from experiments with wood pellet operations

Avg. Temp, T	Fuel feed, FF	Relox, RE	MC	CO	H_2	CO_2	CH_4
T2-T6	tDS/h.m^2			%	%	%	%
700.6	1.07	0.36	0.20	8.14	5.44	15.7	0.98
693.6	0.93	0.42	0.23	5.58	4.5	16.8	1.42
693.6	1.02	0.38	0.35	7.43	5.78	19.7	1.32
718.2	1.02	0.38	0.35	6.2	2.3	15.3	1.1
754.4	1.10	0.35	0.33	9.13	6.1	17.5	1.75
681.8	0.79	0.49	0.41	4.67	3.51	17.2	0.61
737	1.02	0.38	0.40	14.1	3.56	15.5	3.06
774	0.85	0.46	0.44	9.75	8.51	17.2	1.99
800.2	0.68	0.57	0.50	10.3	8.51	18.5	1.76
798.8	0.68	0.57	0.50	9.75	6.55	17.6	1.5
785.8	1.01	0.38	0.40	8.82	6.38	17.4	1.31
809.2	1.19	0.33	0.44	7.11	4.81	16.4	1.1
746.6	1.16	0.33	0.45	7.11	4.2	17.6	1.39
742.6	1.11	0.35	0.46	8.8	4.02	15.6	2.77
762.2	1.36	0.28	0.41	11.6	4.02	14.2	3.2
707.4	1.53	0.25	0.38	11.4	7.33	17.1	3.22
741.4	0.85	0.46	0.44	12.1	5.4	12.1	3.36
702.8	1.18	0.33	0.36	9.81	7.52	16	2.19
671.6	1.10	0.35	0	12.5	6.47	15	3.96

In this case, approximately 4.8 m^3 of air is needed for the 100 % oxidation of 1 kg of wood pellets. The moisture content means the moisture including the steam added. The gas composition also includes N_2 and H_2O which is not presented in the table. The amount of tars and higher CHx is excluded since the content is at the detection level, meaning not very accurate values.

3 Description of Simulation Model Combining Energy and Material Balances with PLS-Models

The simulation model, used in this paper, is developed in Modelica that can be run in both Dymola and Open Modelica. The model is expressed as a semi-steady state model giving the heat and mass balance of the gasification system. The model consists of a heat and mass balances where material flows as well as molar flows of both organics and inorganics are followed through the system with the gasifier, cyclone, G-valve, heat exchanger/cooler and the scrubber.

The gas composition is given by the PLS-models, determined from experimental measurements, for each of the gas components H_2, CO, CO_2 and CH_4, while N_2 is assumed as a ballast from the air fed to the system. H_2O is given by the shift reaction at given conditions and the heat balance of the system based on the partial combustion and the heat losses.

The heating value given for the wood pellets is 18.5 MJ/kg which corresponds to the formula $CH_2O_{1.2}$ at 8% moisture (measured). For pellets made of wood of spruce and pine, the formula $CH_{1.44}O_{0.66}$ (according to the elemental analysis) is expected which corresponds to a higher heating value of 23.7 MJ/kg and a lower heating value of 21.8 MJ/kg assuming 8% moisture.

Table 2. Polynoms (PLS) for prediction of gas composition as a function of operating conditions with respect to average temperature in the reactor, load, relative oxidation and moisture content

Vol %	B_0	A_1	A_2	Relox	MC %	R^2	Q^2
CO	-12.6	0.012	8.46	-0.001	0.07	0.8	0.61
H_2	-13.1	0.017	-0.001	0.0472	0.07	0.8	0.57
CH_4	-6.7	0.002	4.20	0.0392	0.03	0.8	0.69
CO_2	7.2	0.020	0.00097	-0.052	-0.2	0.7	0.7

According to the actual chemical analysis of the wood pellets, $CH_{1.46}O_{0.625}$, the 100 % oxidation demand of air would in this case be 4.8 nm^3/kg DS, for the two compositions. In the calculation for 100% relox in this study the latter value has been used for the wood pellets. All five measured major gas components are correlated to the relox, capacity, MC and the average temperature in the CFB using PLS regression. The results are shown in Table 2. The gas composition, C (vol.%) is calculated for these six major gas components with a polynom as shown in

$$C = B_0 + A_1.T + A_2.FF + A_3.RE + A_4.MC \quad (1)$$

where T is the value of the average temperature in o C of the five positions in the reactor, FF is the load in ton dry solids per m^2 of the reactor cross area per hour, RE is the relative oxidation as the percentage of oxygen in relation to what is needed for 100% oxidation of all organic material and MC is the moisture content in % of the total fuel weight including the added steam. B_0, A_1, A_2 , A_3 and A_4 are the regression constants given in the Table 2.

R^2 is 1.0 when perfect fit of all experimental data into the model, while 0.5 is a value that is a minimum for being able to start using the correlation. Q^2 is the corresponding prediction power when the model is used to predict performance at any condition covered by the experiments. Above 0.5 we can start using the prediction model and the prediction is perfect at 1.0. Here we get 0.6-0.7 for most of the gas components, which makes the models usable, although quite a bit from very good. The moisture (H_2O) is calculated from the shift reaction with the constant KT given for the average temperature (T), assuming we have steady state conditions as shown in

$$K_T = [CO][H_2O]/[H_2]\,[CO_2] \quad (2)$$
$$K_T = (Temp\text{-}649) * (0.154/55) + 0.50800 \quad (3)$$

This is for the actual gasification. For the moisture content in the gas after the scrubber, the water content of the gas at saturation for the given scrubber temperature is used. From this we recalculate the gas composition used in the simulations later on as a function of operating conditions, but then combining also with energy and mass balances. The results from the combined model for wood pellets are presented in Table 3. The composition of wood pellets is assumed as $CH_{1.44}O_{0.66}$ with Mw of 24. We need to add net 42.9 mol O_2/kg fuel and 3.76 mol N_2 per mol O_2 added as air for 100% oxidation.

The relative oxidation and the load has strong impact while the temperature has less effect. However, the steam has no significant impact, i.e. when increasing by 40% from a relatively high level. In the reactor, it is seen that the heat demand for driving the processes varies considerably, i.e. from 30% to about 50%.

It is interesting to note that we cannot detect any tars in the synthesis gas, neither directly at the fuel injection point (2 meters above the injection point) or in the exhaust gas channel before the filter.

Table 3. Results from simulation using the combined modelica and PLS models for gas after condensation to 40°C for wood pellets. MC is moisture content in pellet + steam

Load	Relox	MC	T	CO	H_2	CH_4	CO_2	N_2	H_2O	HHV
$t/m^2.h$	%	%	°C	%	%	%	%	%	%	kJ/m^3
1	35	30	700	8.4	20	1.8	17.6	37.6	14.6	3918
1	25	30	700	9.5	22.7	1.4	20.6	30.5	15.2	4221
1	45	30	700	7.5	17.8	2	15.2	43.2	14.2	3681
1	35	45	700	10.5	20.2	2.5	15.1	38	13.7	4498
1	35	30	800	8.5	19.7	1.8	17.2	37.1	15.8	3898
2	35	30	700	12.4	20.4	4.6	11.4	38.4	12.9	5565
2	25	30	700	14	23.3	4.8	13.4	31.3	13.2	6191
2	45	30	700	11	18.2	4.4	9.8	43.9	12.7	5078
2	35	45	700	13.8	20.5	5.2	9.4	38.5	12.6	5993
2	35	30	800	12	20.2	4.3	11.8	38.1	13.6	5398

The bed weight is ca 21 kg. 2* 0.4 kg is passing over to the filter and the filter ash has 7% respectively 6.2 % C in the ash. The carbon particles are small balls and when the bed material is emptied, the C-content is found to be 1.2 %. The total carbon in the feed is 62 kg and in the residue, is 0.3 kg. The C-conversion is (1- 0.3/62) = 99.5%. We have now been running the experiments five times and tried to control the process in order to verify that no detectable tars are formed and now feel quite convinced that this is a fact. We also have ideas about the reasons for no tar in the gas and will perform tests in future to confirm this experimentally.

4 System Studies

The studies on H_2-production in a CHP plant was presented (Naqvi et al. 2016, 2017). Yang and Ogden (2007) made an overview of production costs for Hydrogen production. Further studies were made on black liquor gasification systems where different cycles and solutions were compared, including among others CO_2 removal (Naqvi et al. 2010, 2017). Asadullah (2014) has made a critical review of down- stream gas cleaning after biomass separation, which includes also particle and tar removal. Concerning the gas separation, H_2 is a very small molecule and thus passes through even tight membranes quite easily compared to most other molecules except water.

By condensing water before the membrane unit, we thus can get relatively pure H_2 in the permeate. The separation between N_2 and H_2 has been commercialized since long ago in the ammonia production. Here relatively large pressure difference between feed and permeate has been applied. Now new membranes are coming where the pressure difference might be only one or a few bars, which makes it easier from a system perspective. Example of such membranes are porous graphene (Du et al., 2011) and PDMS composites with SiO_2 and B_2O_3 (Lee et al., 2015). If we can let H_2 pass through while the rest are remaining in the reject, these can be combusted in a boiler or even a gas turbine with an external combustion chamber. If we want to upgrade the gas further CO_2 is the second easiest gas to separate as it is quite polar and thereby can be dissolved in liquids like MEA (mono ethanol amine) or alkaline solutions like the scrubber solution used to separate H_2S in the black liquor process, but with a higher pH. With pH 11.5 all CO_2 is absorbed. With the MEA it is easy to regenerate the liquid by just heating and stripping off CO_2 again. Using alkaline solution, the $Na_2O.TiO_2$ regeneration can be used, but a bit more complex system solution is needed. The separation with MEA using micro porous membranes was demonstrated by Yuexia et al. (2010) and Lv et al. (2012). Here modification of the polypropylene membrane was made to get long term stabile performance.

In the solution presented in Figure 2, we have a gasification process that is used for the production CH_4. The synthesis gas could be separated to extract methane, while the residual gas could be combusted directly in a boiler, or in an external gas turbine combustor, making a combined cycle possible. The heat from the steam turbine condenser then could be used for the district heating. Even CO_2 can be removed at the far end of the exhaust gas train.

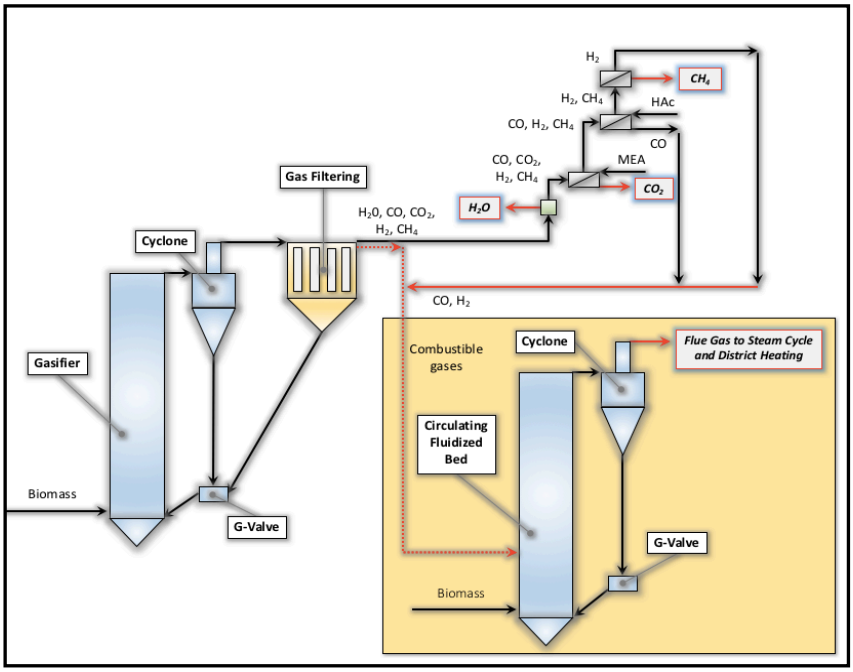


Figure 2. Different bio-refinery systems can be either integrated or operated separately

In a system for black liquor gasification, an alternative with a combi-cycle turned out to be able to give an electric to fuel heating value efficiency of up to 38%, which is high for a process with such a poor fuel (Dahlquist, Jones 2005). If we instead look at alternatives with biomass in a CHP plant, the district heating is very interesting from an energy efficiency perspective but with an issue that the heat demand varies significantly over the year. It is thus interesting to be flexible such as the heat demand must be fulfilled when it is very cold, and then chemicals or electric power may be of secondary importance. As the conditions will vary much more than previously, the dynamics of the system is of interest in order to perform transition from one operation mode to another in a smooth way. This is another important aspect needing the dynamic simulation models like the one in Modelica, and not only steady state models.

In the future, we see a stronger demand to use cheaper fuels in the CHP plants, as the cost for biomass fuels like pellets becomes too high to be competitive in many cases. Organic waste then is interesting as the cost is very low or you even get paid to take care of the waste. By gasifying the waste, remove the particles like alkali salts and corrosive substances like HCl, we can produce clean fuel that can be used in both gas turbines and boilers that are not designed for wastes. The waste has not been directly tested in the pilot plant but should be principally easier than the black liquors that we actually have tested and found feasible to use as fuel.

Concerning the gas separation using membrane filtration, the plan is to test the separating gas compositions that are produced from the gasifier, to determine the efficiency for the mixtures. There are new types of membranes are developed all which gives a high probability for finding suitable membranes to use in the future, and also to give input to membrane developers what to aim for? The regeneration of absorbents used in the liquid-membrane-combinations is another task to develop further. Here the possibility to recover CO_2 and deposit or use it for other purposes is of high potential interest to reduce CO_2 emissions to the atmosphere. This also should have a positive economic impact in the future.

5 Conclusions

In the paper, we have shown how regression models like PLS, PCA and similar can be made from experiments and combined with the dynamic physical models developed in the Modelica. These models can be used to study different systems from the energy and material balance perspective, but also to investigate how to switch from one process mode to another in a smooth way. This has significance as the economic conditions will vary considerably in the future from one time of the day to another, as well as over the season, making it much more complex to fulfil different demands. When there is focus on the conversion processes such as the gasification, we will see an increasing demand for the gas separation, like membrane separation, for developing efficient system solutions. The new demands like CO_2 removal may give different economical optima, if CO_2 is valued significantly higher than today. This will also shift the use of fossil fuels for production of chemicals into a demand to use biomass, which will give new incentives to the proposed processes for production of base chemicals like CH_4 and H_2.

Acknowledgements

We thank Bioregional and especially Sue Riddlestone, for making their pilot plant available for the tests with wood pellets, and Swedish Energy Agency and KKS are acknowledged for the financial support.

References

A. Gómez-Barea and B. Leckner. Modeling of biomass gasification in fluidized bed, *Progress in Energy and Combustion Science*, 36(4): 444–509, 2010.

C. D. Blasi. Kinetic modeling of biomass gasification and combustion, Intelligent Energy Europe (PyNe). (downloaded 5 Jan, 2016).

C. Roberto and M. D. Veroli. Mathematical Modelling of Biomass Gasification in a Circulating Fluidized Bed CFB Reactor, *Journal of Sustainable Bioenergy Systems*, 2: 160-169, 2012.

C. Yang, and J. Ogden. Determining the lowest-cost hydrogen delivery mode, *International Journal of Hydrogen Energy*, 32: 268-286, 2007.

D. Baruah and D. C. Baruah. Modeling of biomass gasification: A review, *Renewable and Sustainable Energy Reviews*, 39: 806–815, 2014.

E. Dahlquist and A. Jones. Presentation of a dry black liquor gasification process with direct caustization, *TAPPI Journal*: 15-19, 2005.

H. Du, J. Li, J. Zhang, G. Su, X. Li, and Y. Zhao. Separation of Hydrogen and Nitrogen Gases with Porous Graphene Membrane, *The Journal of Physical Chemistry*, 11: 23261–23266, 2011.

L. Yuexia, X. Yua, S. Tu, J. Yan, and E. Dahlquist. Wetting of polypropylene hollow fiber membrane contactors, *Journal of Membrane Science*, 362: 444–452, 2010.

M. Asadullah. Biomass gasification gas cleaning for downstream applications: A comparative critical review, *Renewable and Sustainable Energy Reviews*, 40: 118-132, 2014.

M. Naqvi, E. Dahlquist, and J. Yan. Complementing existing CHP plants using biomass for production of hydrogen and burning the residual gas in a CHP boiler, *Biofuels*, 8(6): 675-683, 2017.

M. Naqvi, J. Yan, M. Danish, U. Farooq, and S. Lu. An experimental study on hydrogen enriched gas with reduced tar formation using pre-treated olivine in dual bed steam gasification of mixed biomass compost, In *International Journal of Hydrogen Energy*, 41(25): 10608-10618, 2016.

M. Naqvi, J. Yan, and E. Dahlquist. Synthetic natural gas (SNG) production at pulp mills from a circulating fluidized bed black liquor gasification process with direct causticization, *Conference proceedings of ECOS 2010*.

M. Naqvi, J. Yan, E. Dahlquist, and S. Naqvi. Off-grid electricity generation using mixed biomass compost: A scenario-based study with sensitivity analysis, *Applied Energy*, 201: 363-370, 2017.

S. H. Lee and H. K. Lee. Separation of Hydrogen-Nitrogen Gases by PDMS-$SiO_2B_2O_3$ Composite Membranes, *Membrane Journal* 25 (2), 2015.

Y. Lv, X. Yu, J. Jia, S. Tu, J. Yan, and E. Dahlquist. Fabrication and characterization of superhydrophobic polypropylene hollow fiber membranes for carbon dioxide absorption, *Applied Energy*, 90(1): 167-174, 2012.

Initial Results of Adiabatic Compressed Air Energy Storage (CAES) Dynamic Process Model

Tomi Thomasson Matti Tähtinen

VTT Technical Research Centre of Finland Ltd., Finland, tomi.thomasson@vtt.fi

Abstract

The amount of wind and solar generation has seen exponential growth during the recent decades, and the trend is to continue with an increased pace. Due to the intermittency of the resources, a threat is posed on grid stability and a need created for regulation. One solution to control the imbalance between supply and demand is to store the electricity temporarily, which in this paper was addressed by implementing a dynamic model of adiabatic compressed air energy storage (CAES) with Apros dynamic simulation software. Based on the literature review, the existing models due to their simplifications do not allow transient situations e.g. start-ups to be studied, and despite of its importance, part-load operation has not been studied with sufficient precision. The implemented model was validated against analytical calculations (nominal load) and literature information (part-load), showing considerable correlation. By incorporating the system with wind generation and electricity demand, the grid operation of CAES was studied. In order to enable this, the start-up and shutdown sequences based on manufacturer information were approximated in dynamic environment, to the authors' best knowledge, for the first time. The initial results indicate that the modularly designed model offers an accurate framework for numerous studies in the future.

Keywords: energy storage, compressed air energy storage, dynamic simulation, numerical simulation

1 Introduction

The rise of variable renewable energy (VRE) has been remarkably rapid during the past decades; not only it is considered an integral part of the current energy system, its importance in the future cannot be highlighted enough (World Energy Council, 2013). This new generation capacity mainly consisting of solar and wind power lacks inertia and carries the burden of intermittency, meaning that the availability of the resource varies in both short and long term (Barnhart *et al*., 2013). Both solar and wind have seen the installed capacity increasing exponentially during the recent decade, but the share of VRE generation in the electricity grids of the large markets is still considerably modest (REN21, 2015). Due to intermittency, technical barriers e.g. disturbances in voltage magnitude and delivery frequency may create economic barriers in transition to greater share of VRE generation (Sundararagavan and Baker, 2012; SBC Energy Institute, 2013). For example, curtailment of around 40% of the total generation has been reported in the wind farms of China (Christiansen and Murray, 2015).

Solutions to counter the effects of intermittency range from demand side response e.g. time-of-day electricity pricing to utility side response, amongst which the compressed air energy storage (CAES) belongs. The operating principle of CAES is best described as mechanical conversion of electricity into the form of pressurized air. Electricity is stored during the hours of low consumption and supplied back to the grid once the demand has increased. CAES as a process is fundamentally dynamic, as the system is able to ramp twice or three times as fast as the conventional alternatives (Bradshaw, 2000) and is designed to accommodate multiple start-ups and shutdowns on a daily basis (Schulte *et al*., 2012). The two existing CAES systems are based on the needs of the conventional energy system. While the Huntorf CAES was initially constructed to provide black start capability for nuclear power plants (Succar and Williams, 2008), the McIntosh CAES fills the deficit between the capacity of a coal power plant and the demand (Arsie *et al*., 2007).

The deficiency of diabatic CAES technology introduced above is the requirement of an external source of heat, typically natural gas, leading to carbon dioxide emissions. An alternative is to recover, store, and utilize the heat of compression, which is the working principle of the adiabatic configuration (Kreid, 1976). This approach is generally accepted to lead to notably higher cycle efficiency compared to diabatic CAES (Hartmann *et al*., 2012), but despite extensive research activities (Zunft, 2015), the technology is only nearing demonstration stage (Airlight Energy, 2016). In order to overcome challenges mainly related to temperature limitations in the compressor technology (De Biasi, 2009) and storing the thermal energy, low temperature concept has been recently proposed (Wolf and Budt, 2014). As the cycle efficiency is not governed by the Carnot efficiency, storing the heat at a lower temperature level

leads to a relatively small efficiency penalty, but allows faster start-up due to lower thermal inertia (Freund *et al.*, 2012). This, along the economic limitations (Zunft, 2015) and the required flexibility in the electricity market (Wolf and Budt, 2014), supports the idea of developing downscaled systems.

2 Methodology

As CAES is comparable to open Brayton cycle gas turbine, great understanding of the process and its components readily exists. However, the need for detailed dynamic simulations has recently been pointed out (Budt *et al.*, 2016). This chapter presents the background for the model development of this paper.

2.1 The lack of dynamic features in the existing models

Although dynamic as term well describes the operation of CAES, the past efforts in literature have largely focused on the dynamic cavern operation (Nielsen and Leithner, 2009; Khaitan and Raju, 2013). Detailed studies related to turbomachinery have scarcely been conducted, even then typically on component level based on comprehensively studied analytical approaches (Luo *et al.*, 2016). Furthermore, the use of constant material properties limits the accuracy of several of the models; although at times justified, the error in one property is also present in the other.

As the idea of combining energy storage with intermittent generation is well understood, logic systems related to CAES have been presented in literature (Arsie *et al.*, 2007; Zhao *et al.*, 2015). The limitation in steady-state simulation is that the logic systems are often developed by assuming fixed time windows; a disturbance occurs whenever it is selected to take place. In reality, there is a clear need to operate the storage depending on the supply and demand while maintaining the system as efficient as possible. This is only achieved by means of control engineering, which seemingly has only been covered by Budt *et al.* (2012) so far.

2.2 Apros dynamic simulation software

Apros is a dynamic simulation tool and modelling software, which includes tools for design, evaluation and testing of various types of processes (Fortum and VTT). As suggested by Figure 1, Apros enables the possibility for detailed control and operation strategy development besides process simulation. Apros has commonly found applications in power plant simulations, but due to comprehensively validated component model library and the possibility for user-defined components, the software has more recently

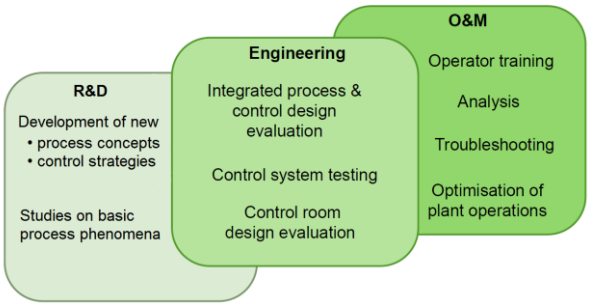


Figure 1. Apros and its various applications.

been successfully extended to the field of energy storage (Tähtinen *et al.*, 2016).

The model hierarchy of Apros consists of three levels: diagrams, process component level and calculation level. The user often operates on the diagram level, constructing the model out of the process components such as pipes, valves and heat exchangers; each operating analogous to their concrete counterpart. When included in the model, the process component automatically creates its necessary calculation level objects, the nodes and branches. With these objects, the conservation equations for mass, momentum and energy are solved. Several accuracy levels for solving the thermohydraulic solution are available; for the composition AIR used in the dynamic model, representing ideal gas mixture of oxygen and nitrogen, only the homogenous flow model may be selected.

2.3 Implementation of the dynamic model

The dynamic model is developed according to the low-temperature principles introduced by Wolf and Budt, (2014) leading to a system with a relatively low compression and expansion power. Most importantly, the goal is to create a flexible model, allowing easy scalability with future studies in mind. In order to validate the model, the initial conditions are based on thermodynamic analysis, for which the resulting parameters are shown in Table 1. Values for certain constants such as polytropic efficiencies and nominal storage pressure are obtained from literature (De Biasi, 2009; Barbour *et al.*, 2015). The layout of the model, simplified in Figure 2, can be considered to consist of the following subsystems:

- Turbomachinery
- Thermal energy storage (TES)
- Compressed air storage (CAS)
- Control and logic system

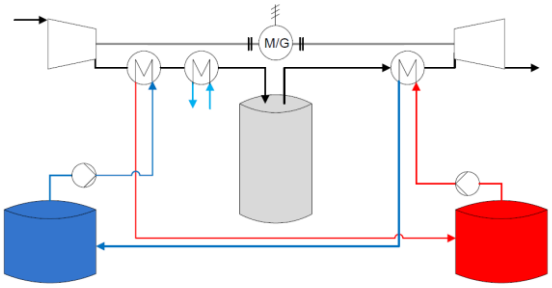


Figure 2. Simplified illustration of the model layout and the flow schemes: air (black), thermal oil (blue and red) and water (light blue).

As fixed-speed compressor trains have prevalently been used with CAES, variable guide vanes (VGV) have generally been selected for compressor capacity control (Dresser-Rand, 2015). In principle, the guide vanes allow load changes by shifting the operating point. After evaluation, the compressor map presented by Brasz (1996) was selected and introduced to the model as nominal polytropic efficiency and nominal mass flow rate as a function of guide vane angle. Each of the compressors is during periods of low mass flow rates effectively protected from surge by control principles presented by Brun and Nored (2008).

The expander train of CAES can be operated with either fixed or varying inlet pressure (Weber, 1975). Similar to the existing systems, throttle valve placed upstream the expander train is selected to reduce the pressure of the air discharged from the storage, allowing capacity control with an increased efficiency at part-load. The turbomachinery is selected to be connected on a single shaft and to share the electric motor-generator unit, preventing charging and discharging from taking place simultaneously. The clutches, enabling the required loading and unloading of the electric machine, are implemented by using a number of switch and set dominant latch components.

The TES is implemented based on the previous work of Hakkarainen and Tähtinen (2016), in principle solving the mass and energy balances in calculation level. Due to the approach, the tanks are assumed isothermal and only single-phase fluids can be realistically evaluated. Therminol VP-1 (Solutia, 1999) due to its suitable temperature range is selected as the heat transfer fluid. The heat exchangers are dimensioned with information about typical configurations and operating parameters (Freund and Moreau, 2012). During the charging process, heat considered as excess is left to the process. To overcome this issue, auxiliary heat exchanger applied by many including (De Biasi, 2009) is placed upstream to CAS. For the CAS representing an artificial storage tank, heat losses based on conduction and convection mechanisms are implemented.

The combined control and logic system of the model has four primary tasks: to control both temperature and power, to actuate sequences, and to schedule the storage operation. Both the control tasks are implemented by using PI controllers, for which cascade loops are preferred due to smoother control action. For the power regulation, a limit is imposed on the controller set point gradient, which enables selecting the ramp rate based on actual concepts. The role of the sequences is to activate the individual control systems at the correct time during the start-up and shutdown, which are replicated using manufacturer information (Dresser-Rand, 2015). In addition, the storage operation is scheduled by using an interlocked predictive-reactive logic system. By applying a set of boundary conditions, the logic system allows the storage to be only operated at times when certain conditions e.g. deficiency in generation and sufficient storage pressure are met.

Table 1. Key input parameters of the model.

Parameter	Value
General	
Nominal charging power	25.41 MW
Nominal discharging power	14.96 MW
Nominal cycle efficiency	58.87%
Charging period ratio	1.0
Compressor train	
Number of stages	4
Nominal mass flow rate	35 kg/s
Nominal pressure ratio, stage 1	4.33
Nominal pressure ratio, stage 2–4	3.4
Polytropic efficiency	85%
Expander train	
Number of stages	3
Nominal mass flow rate	35 kg/s
Nominal pressure ratio, stages 1–3	4.87
Polytropic efficiency	85%
TES	
Nominal mass flow rate	varies per stage
Nominal heat transfer effectiveness	75 – 85 %
Nominal cold tank temperature	25ºC
Nominal hot tank temperature	188.8ºC
Storage capacity	4 h
CAS	
Nominal temperature	40ºC
Nominal pressure	155 bar
Throttle pressure	120 bar
Storage capacity	4 h

2.4 Setup of the VRE framework

As the existing CAES systems are based on predictable demand of storage operation, there is a need to develop more flexible storage operation strategies. For example, the daily and weekly cycles presented by Goodson (1992) are by no means viable for VRE. Furthermore, as the entire control system can only be effectively studied when combined with VRE, a simple

wind farm model based on the fundamental equation of wind power (Manwell *et al*., 2009) and commercial wind turbine design (Vestas Wind Systems, 2004) is implemented in Apros.

The validated wind turbine model receives the wind data retrieved from NREL (2015) at a temporal resolution of one minute as an input through a gradient. Realistic representation of hourly load variation is retrieved from Fingrid (2015). As scaling of the storage system according the load curve is important, the load data is scaled to match the nominal output of the system accordingly (Le *et al*., 2012). In addition, the wind data is scaled in order to evaluate each part of the control system and the transitions between the operation modes.

3 Results

The challenge in model validation is the lack of available reference data. One should hence consider the observation of trends and physical phenomena more important than the absolute results. In Figure 3 and Figure 4, the values are consequently presented as relative expression; for the model as a comparison to nominal value and for the reference data as a comparison to maximum value.

Due to the implemented VGV component, the operation of the compressor train is first validated at the design point against the results of the thermodynamic analysis. The simulated power consumption and discharge temperature vary from the reference values for less than one percent for each compressor. At part-load, the created quadratic least squares polynomial fits at maximum create a relative error of 5% in both mass flow rate and polytropic efficiency. In addition, the start-up sequences are validated against the data of Dresser-Rand (2015), showing excellent correlation.

3.1 Qualitative validation of the compressor train

In order to validate the performance of compressor train, the results are compared to those of Budt *et al.* (2012) over the charging period. It should be noted that the reference model (*ref*) consists of eight compression stages, for which the detailed input parameters are not presented by the authors. The comparison is therefore done between the first and the last of the stages, and in addition, the fourth and the sixth stage of the reference model are visualized. In the developed model (*sim*), the compressor power consumption set point is fixed to 20 MW and the storage is charged for approximately four hours from 120 bars to 155 bars.

Figure 3 shows the comparison of compressor power consumption, from which two similarly shaped groups of curves can be observed. While the total power consumption remains constant, the relative power consumption of the first stages decreases. Simultaneously, the last stages show the greatest increase in relative power consumption, as indicated by

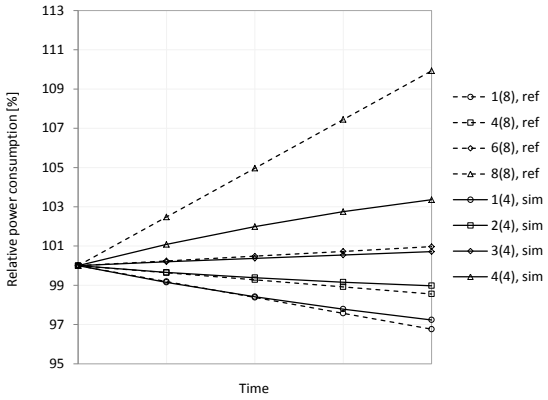


Figure 3. Simulated compressor power consumption compared to the results of Budt *et al*. (2012) over charging period.

both simulations. This is expected as the last stages also are subject to the relatively greatest increase in pressure ratio and hence are forced to operate farthest from the nominal point.

3.2 Qualitative validation of the expander train

The validity of the expander train is analyzed by comparing the relative pressure ratio at varied load rate against the results of Zhao *et al*. (2016) as shown in Figure 4. The authors study a system with two expansion stages, and explain that the decrease in pressure ratio is caused by the relationship between Stodola's ellipse law (Cooke, 1985) and mass flow rate. As the only constant discharge pressure in the expander train is the final point representing ambient conditions, the pressure ratio of the last stage should decrease with load rate, which is confirmed by both the simulations. However, with fixed inlet pressure operation mode, both the inlet and discharge pressure of the first stage decrease when operating at part-load. Therefore, the pressure ratio should largely stay unaffected by the variation in load rate, which is not shown by the results of Zhao *et al*. (2016).

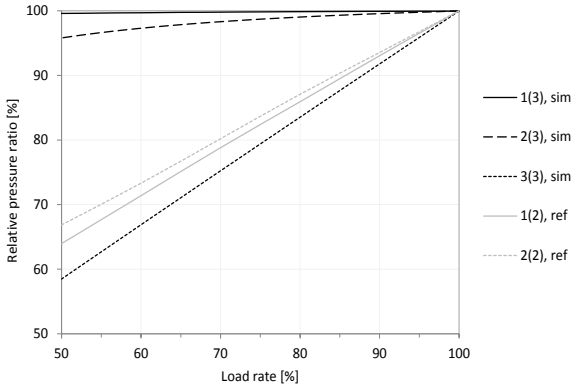


Figure 4. Simulated compressor power consumption compared to Zhao *et al*. (2016) over charging period.

3.3 Dynamic operation of the model

In order to demonstrate the control system and its ability to regulate the power consumption and generation according the demand, in Figure 5 the model is operated for six hours in dynamic conditions. The upper subfigure shows the power charged to or from the grid depending on the excess or deficiency in wind generation, while the lower subfigure illustrates the state of CAS and TES. The following operation is primarily expected from the model during each of the temporal segments:

- Hour 1: charging
- Hour 2: no operation
- Hour 3: charging
- Hour 4: no operation
- Hour 5: discharging
- Hour 6: discharging

The results indicate that the expected operation is to large extent fulfilled. For example, during the first hour the lower subfigure of Figure 5 shows the CAS pressure steadily increasing, while thermal oil is transferred from the cold TES tank to the hot TES tank at elevated temperature level. The flat plateaus in the lower subfigure indicate that the system is not active, whereas the predominant flatness of the area curve in the upper subfigure suggests that the system is to large extent operated with nominal load rate.

The capability for accurate load following is highlighted particularly after t = 2 h, where the system has to adjust the compressor power consumption nearly continuously. Furthermore, even though the operation is only scheduled by using ten-minute-ahead wind forecasts, the signals for unnecessary start-ups and shutdowns are largely avoided. Exceptions of this are visible for example slightly before t = 3 h, where the system is needlessly shut down and started up multiple times consecutively. If the operation was to be scheduled optimally, the problem could be solved with additional interlock mechanisms e.g. inclusion of longer-term forecasts.

4 Conclusions

The initial results show that a dynamic model of adiabatic CAES has been successfully implemented with proven Apros dynamic simulation software. Due to a variety of readily validated process components as well as analytically developed and validated user-defined inclusions, both the steady-state and transient phenomena related to CAES can be studied in more detail than before.

Despite the simple model structure, more advanced phenomena e.g. compressor surge can be readily studied in future research. Model predictive control as well as improved forecasts and economic boundary conditions can be easily included in the existing logic system in order to enable more comprehensive case studies. The current TES model does not include heat losses, which currently limits the possibility for extended simulations. This, however, is to be corrected with the upcoming version of Apros, also allowing more accurate implementation of user-defined heat transfer fluids.

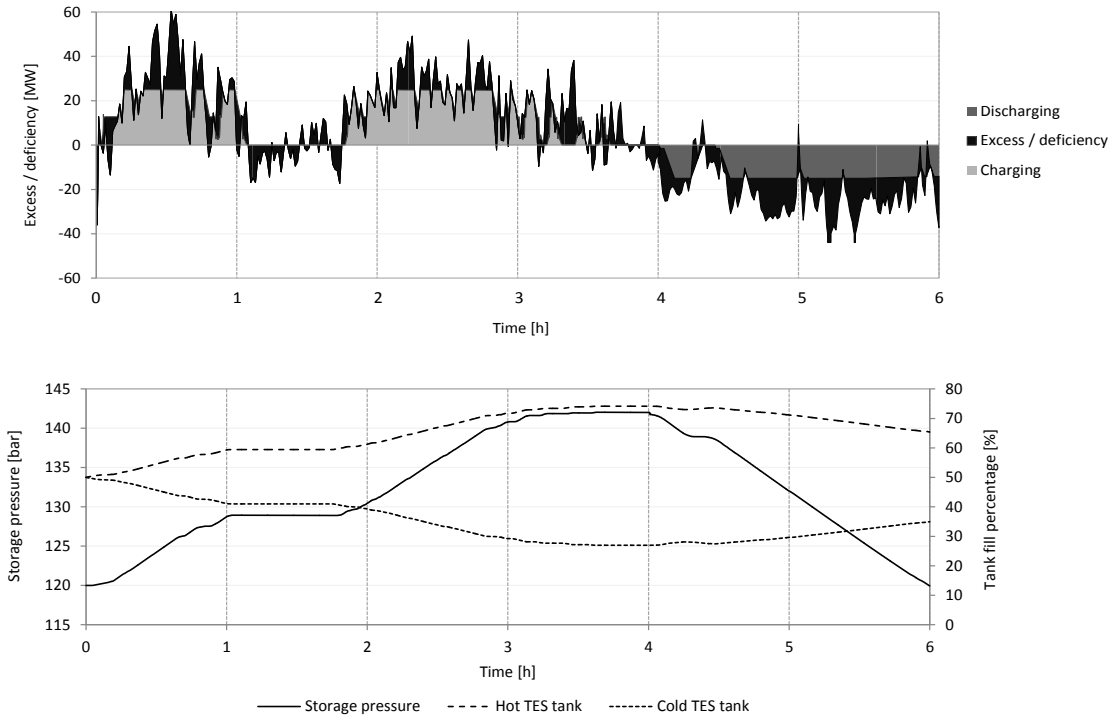


Figure 5. Grid operation of the dynamic model (upper subfigure) and the consequent variation in the state of CAS and TES.

References

Airlight Energy. *Advanced adiabatic compressed air energy storage* [Online], 2016. http://www.airlightenergy.com/advanced-adiabatic-compressed-air-energy-storage/.

I. Arsie, V. Marano, M. Moran, G. Rizzo, and G. Savino. Optimal Management of a Wind/CAES Power Plant by Means of Neural Network Wind Speed Forecast. In *European Wind Energy Conf. and Expo*., 2007.

E. Barbour, D. Mignard, Y. Ding, and Y. Li. Adiabatic Compressed Air Energy Storage with packed bed thermal energy storage. *Applied Energy* 155: 804–815, 2015.

C. Barnhart, M. Dale, A. Brandt, and S. Benson. The energetic implications of curtailing versus storing solar- and wind-generated electricity. *Energy & Environmental Science* 6(10): 2804–2810, 2013.

D. Bradshaw. Pumped hydroelectric storage (PHS) and compressed air energy storage (CAES). In *Power Engineer Society Summer Meeting*, 2000.

J. Brasz. Aerodynamics of Rotatable Inlet Guide Vanes for Centrifugal Compressors. In *International Compressor Eng. Conf.*: 761–766, 1996.

K. Brun and M. Nored. *Application guideline for centrifugal compressor surge control systems* [Online], 2008. http://www.gmrc.org/documents/GMRCSurgeGuideline_000.pdf.

M. Budt, D. Wolf, R. Span, and J. Yan. A review on compressed air energy storage: Basic principles, past milestones and recent developments. *Applied Energy* 170: 250–268, 2016.

M. Budt, D. Wolf, and R. Span, Modeling a Low-temperature Compressed Air Energy Storage with Modelica. In *Proc. 9th Int. Modelica Conf*.: 791–800, 2012.

C. Christiansen and M. Murray. *Energy Storage Study* [Online], 2015. http://arena.gov.au/files/2015/07/AECOM-Energy-Storage-Study.pdf.

D. Cooke. On Prediction of Off-Design Multistage Turbine Pressures by Stodola's Ellipse. *J. Eng. Gas Turbines Power* 107(3): 596–606, 1985.

V. De Biasi. Fundamental analyses to optimize adiabatic CAES plant efficiencies. *Gas Turbine World* 39(5): 2009.

Dresser-Rand. *Dresser-Rand SMARTCAES – Compressed Air Energy Storage Solutions: Enabling Renewables* [Online], 2015. https://www.dresser-rand.com/wp-content/uploads/2015/01/caes.pdf.

Fingrid. *Load and generation* [Online], 2015. http://www.fingrid.fi/en/electricity-market/load-and-generation/.

Fortum and VTT. *Apros* [Online]. http://www.apros.fi/filebank/193-Apros_Combustion.pdf.

S. Freund and R. Moreau. Commercial Concepts for Adiabatic Compressed Air Energy Storage. In *7th International Renewable Energy Conf. (IRES)*, 2012.

S. Freund, R. Schainker, and R. Moreau. Commercial Concepts for Adiabatic Compressed Air Energy Storage. In *7th International Renewable Energy Conf. (IRES)*, 2012.

J. Goodson, *History of first US compressed air energy storage (CAES) plant (110-MW–26 h)* [Online], 1992. http://www.osti.gov/scitech/biblio/6843143.

E. Hakkarainen and M. Tähtinen. Dynamic Modelling and Simulation of Linear Fresnel Solar Field based on Molten Salt Heat Transfer Fluid. *AIP Conf. Proc*. 1734: 070014, 2016.

N. Hartmann, O. Vöhringer, C. Kruck, and L. Eltrop. Simulation and analysis of different adiabatic Compressed Air Energy Storage plant configurations. *Applied Energy* 93: 541–548, 2012.

S. Khaitan and M. Raju. Dynamic simulation of air-storage based gas turbine plants. *International Journal of Energy Research* 37(6): 558–569, 2013.

D. Kreid. *Technical and economic feasibility analysis of the no-fuel compressed air energy storage concept* [Online], 1976. http://www.osti.gov/scitech/biblio/7153562.

H. Le, S. Santoso, and T. Nguyen. Augmenting Wind Power Penetration and Grid Voltage Stability Limits Using ESS: Application Design, Sizing, and a Case Study. *IEEE Trans. Power Syst*. 27(1): 161–171, 2012.

X. Luo, J. Wang, C. Krupke, Y. Wang, Y. Sheng, J. Li, Y. Xu, D. Wang, S. Miao, and H. Chen. Modelling study, efficiency analysis and optimisation of large-scale Adiabatic Compressed Air Energy Storage systems with low-temperature thermal storage. *Applied Energy* 162: 589–600, 2016.

J. Manwell, J. McGowan, and A. Rogers. Wind characteristics and Resources. In *Wind Energy Explained*, 2nd ed. Chichester, United Kingdom: Wiley, 23–90, 2009.

L. Nielsen and R. Leithner. Dynamic Simulation of an Innovative Compressed Air Energy Storage Plant - Detailed Modeling of the Storage Cavern. *WSEAS Trans. Pow. Syst*. 4(8): 253–263, 2009.

NREL. *National Wind Technology Center: NWTC M2 Tower* [Online], 2015. http://www.nrel.gov/midc/nwtc_m2/.

REN21. *Renewables 2015 Global Status Report* [Online], 2015. http://www.ren21.net/wp-content/uploads/2015/07/REN12-GSR2015_Onlinebook_low1.pdf.

SBC Energy Institute. *Electricity Storage* [Online], 2013. https://www.sbc.slb.com/~/media/Files/SBC%20Energy%20Institute/SBC%%20Institute_Electricity_Storage%20Factbook_vf1.pdf.

R. Schulte, N. Critelli, K. Holst, and G. Huff. *Lessons from Iowa: Development of a 270 Megawatt Compressed Air Energy Storage Project in Midwest Independent System Operator* [Online], 2012. http://www.sandia.gov/ess/publications/120388.pdf.

Solutia. *Therminol VP-1* [Online], 1999. http://www.sintelub.com/files/therminol_vp1.pdf.

S. Succar and R. Williams. *Compressed Air Energy Storage: Theory, Resources, And Applications For Wind Power* [Online], 2008. https://www.princeton.edu/pei/energy/publications/texts/SuccarWilliams_PEI_CAES_2008April8.pdf.

S. Sundararagavan and E. Baker. Evaluating energy storage technologies for wind power integration. *Solar Energy* 86(9): 2707–2717, 2012.

M. Tähtinen, T. Sihvonen, J. Savolainen, and R. Weiss. Interim H2 storage in Power-to-X Process: Dynamic Unit Process Modelling and Dynamic Simulations. In *10th Int. Renewable Energy Storage Conf. (IRES)*, 2016.

O. Weber. *The Air-Storage Gas Turbine Power Station at Huntorf* [Online], 1975. http://www.nrri.umn.edu/egg/REPORTS/CAES/References/Weber.pdf.

Vestas Wind Systems. *General Specification: V90 – 3.0 MW* [Online], 2004. http://www.gov.pe.ca/photos/sites/envengfor/file/950010R1_V90-GeneralSpecification.pdf.

D. Wolf and M. Budt. LTA-CAES - A low-temperature approach to Adiabatic Compressed Air Energy Storage. *Applied Energy* 125: 158–164, 2014.

World Energy Council. *World Energy Scenarios: Composing energy futures to 2050* [Online], 2013. https://www.worldenergy.org/wp-content/uploads/2013/09/World-Energy-Scenarios_Composing-energy-futures-to-2050_Full-report.pdf.

Modeling of Black Liquor Gasification

Erik Dahlquist[1], Muhammad Naqvi[1], Eva Thorin[1], Jinyue Yan[1], Konstantinos Kyprianidis[1], Philip Hartwell[2]

[1]School of Sustainable Development of Society and Technology, Mälardalen University (MDH), Sweden,
{erik.dahlquist,raza.naqvi, eva.thorin, jinyue.yan,
konstantinos.kyprianidis}@mdh.se
[2]BioRegional MiniMills (UK) Ltd., United Kingdom

Abstract

The energy situation in both process industries and power plants is changing. It is becoming interesting to perform system analysis on how to integrate gasification into chemical recovery systems in the pulp & paper industry and into the CHP systems in power plant applications to complement with production of chemicals aside of heat and power. The potential chemicals are methane, hydrogen, and methanol. It is also interesting to estimate the potential to introduce combined cycles with gas turbines and steam turbines using both black liquors and other type of biomass like pellets, wood chips etc. To perform such type of analysis, it is vital to have relevant input data on what gas composition we can expect from running different types of feedstock. In this paper, we focus on black liquors as feedstock for integrated gasification systems. The experimental results are correlated into partial least squares models to predict major composition of the synthesis gas produced under different conditions. These quality prediction models are then combined with physical models using Modelica for the investigation of dynamic energy and material balances for complete plants. The data can also be used as input to analysis using e.g. ASPEN plus and similar system analysis tools.

Keywords: black liquor, gasification, CHP, Modelica, physical models, synthesis gas

1 Introduction and Literature Review

During the 70' s, there was a strong demand to increase steel production from the iron ore. ASEA together with Stora and LURGI thus started the development of a new process, the circulating fluidized bed gasification (CFBG). The CFBG process was tested at a demo-plant in Vasteras in a 0.5 m inner diameter reactor with 20 m height. LURGI built a number of coal gasification plants and ASEA became ABB and afterwards ABB acquired the American company Combustion Engineering in 1990. Then, black liquor gasification (BLG) became interesting and a pilot plant was operated in Västerås.

Dipal and Baruah (2014) made an overview of biomass gasification modelling recently and different modeling approaches were categorized based on criteria such as type of gasifier, feedstock, modeling considerations and evaluated parameters. Gómez- Barea and Leckner (2010) performed the modeling work performed with many different approaches from artificial neural nets to computational fluid dynamics. They covered conversion of single fuel particles, char, and gas and conclude that most of the different approaches fit quite well between models and experimental results. However, a very little work has been made on real gasifiers or systems at a larger scale. Capata and Veroli (2012) made a mathematical model over an air-blown CFB with a capacity of 100 kWth. They concluded that there were some problems to get reasonable predictions of the tar formation. It is interesting to note that we did not create any detectable amounts of tar at all while running our CFB gasifier (100-200 kWth) with wood pellets. This shows that the gasification results are influenced depending on the fuel and how the plants are operated. It becomes difficult to achieve accurate models correlating to the experiments, if the mechanisms are not completely understood. Blasi (2016) has made an overview of the kinetic processes in detail to describe tar formation from a theoretical perspective. Still, it is interesting to describe what is actually taking place inside the reactors to be able to predict the process.

2 Description of the Pilot Plant

The experimental work has been performed in a CFB gasifiers in Vasteras at ABB and was built on the design developed through the cooperation between ASEA and LURGI during the 70's. The reactor had a diameter of 170 mm and a height of 10 meters, integrated with one cyclone with a G-valve. The synthesis gas was cooled through a heat exchanger and the gas was cleaned in a bag filter first, then in a scrubber, and finally combusted. At the G-valve entrance, the dust was recirculated from the bag filter. The gas sampling was made using a NIR meter giving simultaneous analysis of several gases.

Table 1. Results from experiments with black liquor operations

Avg. Temp	Fuel feed	Relox	MC	CO	H_2	CO_2	CH_4
T2-T6	tDS			%	%	%	%
655	8	0.45	0.35	2	4	13.5	1.2
597	8	0.54	0.35	2.2	7.5	16.1	1.2
686	16	0.29	0.35	4	11.7	15.7	2.1
703	16	0.36	0.35	3.3	13.3	19.1	1.9
705	20	0.35	0.35	3.9	12.2	18.8	2.3
646	11	0.45	0.35	4.1	8.7	16.7	1.6
654	19	0.26	0.30	3.7	15	14.5	1.8
676	17	0.36	0.30	3.9	12.9	14.2	1.1
613	14	0.38	0.3	2.8	12.3	9.5	0.9
678	14	0.36	0.3	2.5	12.4	8.9	1
677	14	0.35	0.3	2.2	9.3	6.9	1.1
674	13	0.52	0.42	2.8	8.3	15	1.2
677	13	0.53	0.42	2.8	8.1	15.8	1.1
611	15	0.36	0.32	3.7	12	15.2	1.4
612	16	0.33	0.32	3.6	12.4	15.1	1.5
678	15	0.34	0.3	3.4	11	6	1.1

The gas sampling was made at several points in the reactor, although the main position was after the filter. The black liquor gasifier was operated for several years and experiments run in accordance with factorial design of the variation of the operating variables: temperature, relative oxidation, capacity, pressure, different black liquors, addition of KCl, operation with TiO_2, recycling of dust from bag house and others.

3 Experimental results

In Table 1, we have presented few results extracted from our large data set. The selections have been made to have representation of the whole operational volume with similar amount of samples for each condition, to get balanced models. Every experimental test has been operated at least four hours under as steady state conditions as possible. The fuel rate is shown in ton DS/h.m^2 based on the reactor size. The 20 kg/h fuel load in the reactor corresponded to 1.13 ton tDS/m^2.h. The relative oxidation (Relox) means the amount of air (m^3) needed for the 100 % oxidation of 1 kg of fuel (dried solids). In this case, approximately 4.9 m^3 of air is needed for the 100 % oxidation of 1 kg of feed. The moisture content means the moisture including the steam added.The amount of tar was found to be very low that could not be determined accurately during steady state operations, although some tar was formed during the start-up phase. Since tar were not found considerably, they are excluded under the synthesis gas compositions. In addition, N_2 and H_2O are also not included in Table 1. It is found that there is a significant difference between the gas composition obtained from black liquor and wood, gasified in principally using the same reactor (Naqvi et al, 2010, 2016, 2017a).

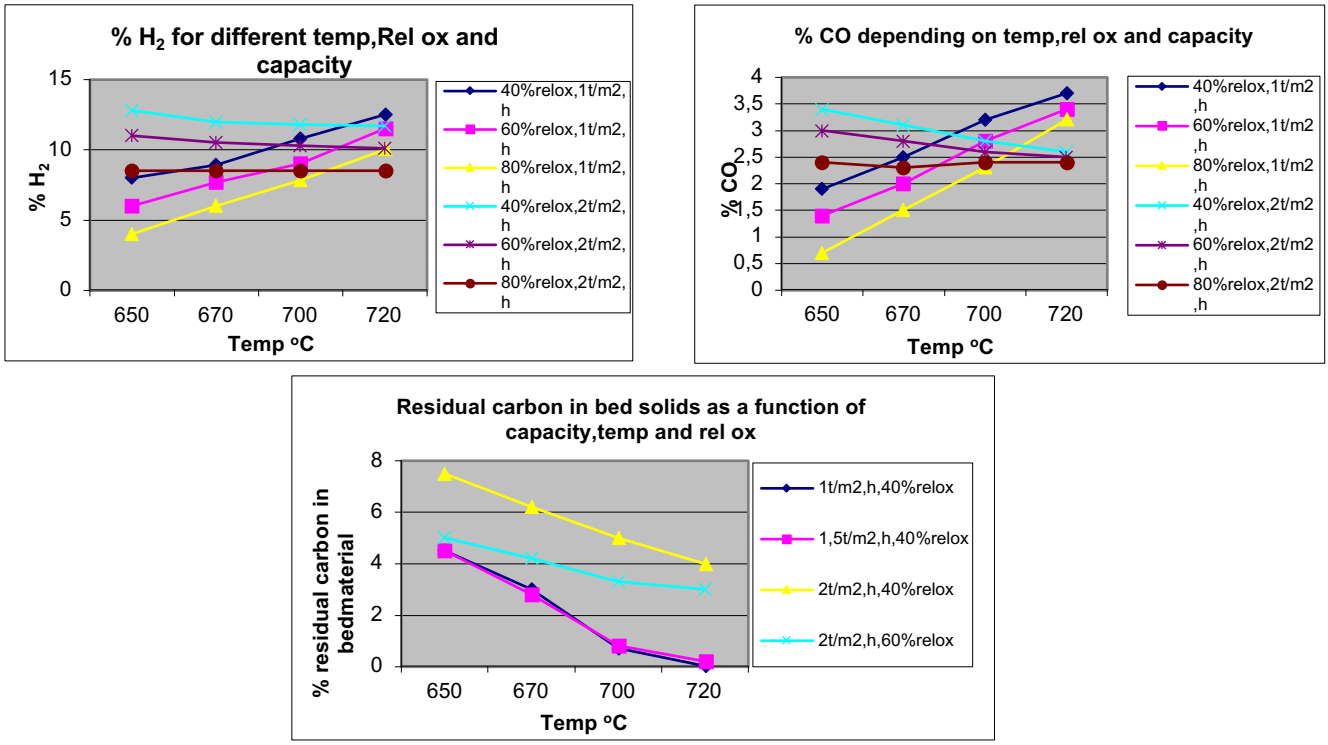


Figure 1. Correlation between H_2 and CO composition as well as residual carbon in fly ash

Table 2. Results from the simulation with black liquor gasification using the combined physical model for energy and material balances and PLS models for gas composition

	EXPERIMENTAL RUNS								
Input	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
DS, %	70	70	70	70	70	70	70	70	70
Feed rate, ton DS/m^2h	1.8	1.8	1.8	**2.4**	**1.2**	1.8	1.8	**2.4**	1.8
Relox, %	35	**45**	**25**	35	35	35	35	35	35
Temp bottom, ^{o}C	700	700	700	700	700	**725**	700	**725**	**725**
Temp at BL-injection, ^{o}C	695	695	695	695	695	720	695	720	720
Temp after scrubber, ^{o}C	65	65	65	65	65	65	**40**	**40**	**40**
Output									
Theoretical possible heat prod, kW	-137	-139	-136	-161	-96	-143	-137	-169	-143
Heat consumed in reactor, kW	44	48	40	58	29	48	44	66	48
Vol. % in wet gas									
H_2O	33	29.4	37.3	34	32.2	35.1	33	39.9	35.1
H_2	10.5	8.7	12.8	9	10.6	13	10.5	9.6	13
CH_4	1.16	0.96	1.43	1.26	1	1.33	1.16	1.33	1.32
CO_2	12	11.2	12.9	10.6	12.9	10.8	12	9.8	10.8
CO	2.5	2.1	2.85	1.8	2.7	2.8	2.5	1.68	2.9
N_2	40.3	46.9	32.1	42.6	40	36.5	40.3	37.2	36.5
H_2S	0.57	0.56	0.6	0.7	0.56	0.43	0.58	0.52	0.43
Heating value dry gas, kJ/kg of BL	-4635	-4284	-5039	-3883	-4633	-6147	-4635	-4668	-6147
Heating value wet gas, kJ/nm^3	-1887	-1580	-2290	-1672	-1874	-2268	-1888	-1753	-2268
Velocity upper reactor, m/s	5.5	6.1	4.9	6.95	3.7	6.2	5.5	8.2	6.2
Flame temp (^{o}C), Air surplus 1.1	1221	1154	1293	1182	1231	1246	1315	1274	1336
Flame temp (^{o}C), Air surplus 1.0	1272	1197	1354	1229	1282	1305	1376	1334	1405
Condensate, mol/kg of BL	9.4	6.4	12.2	9.9	8.8	12.8	25.2	34	30.2
Vol. % in gas after scrubber									
H_2O	26.7	25.5		27	26.4	27.4	13	15.8	13.5
H_2	11.5	9.2		9.95	11.5	14.5	13.6	13.5	17.3
CH_4	1.27	1		1.39	1.1	1.48	1.5	1.87	1.77
CO_2	13.1	11.9		11.7	14	12.1	15.6	13.7	14.4
CO	2.7	2.3		2	2.9	3.2	3.2	2.4	3.8
N_2	44.1	49.5		47.1	43.5	40.8	52.4	52.1	48.7
H_2S	0.63	0.59		0.78	0.61	0.48	0.75	0.73	0.57
Heating value of dry gas, kJ/nm^3	**-2065**	**-1668**		**-1848**	**-2035**	**-2536**	**-2452**	**-2457**	**-3021**
Product gas/air	1.96	1.68		1.85	1.97	2.16	1.96	2.13	
Air, nm^3/h per ton DS/m^2h	3222	4143		4296	2148	3222	3222	4296	
H_2S removal, %			46.9	58.5	50	41.8	50.7	49.7	41.9
SO_4 reduction, %			90.8	86.9	93.3	92.2	91.3	87.9	92.2
C-conversion			**91**	**78**	**98.9**	**97.6**	**92.8**	**82.5**	**97.6**

There are much lower levels of H_2 in wood gasification than BLG but instead higher CO levels. For wood pellets gasification, higher CH_4 concentrations are obtained as compared to black liquor gasification.

The experimental data (not shown here) has also been gathered for the reduction of SO_4 and calculation of the carbon conversion, i.e. balance between what is gasified respective to unconverted carbon in the bed solids dust. In the scrubber, a selective absorption of H_2S takes place while limiting the absorption of CO_2 as far as possible. In the experiments at the pilot plant, we have achieved 1 M (32 g S/l) at a selectivity of 20 by avoiding turbulence in the liquid, but promoting turbulence in the gas phase. In addition, pH is kept as constant as possible at 10.5, to give fast reaction of H_2S. While CO_2 get a back pressure in the liquid film reducing the absorption that is kinetically much slower than the one for H_2S absorption. Few examples of the simulation results are shown in Fig. 1 for the black liquor.

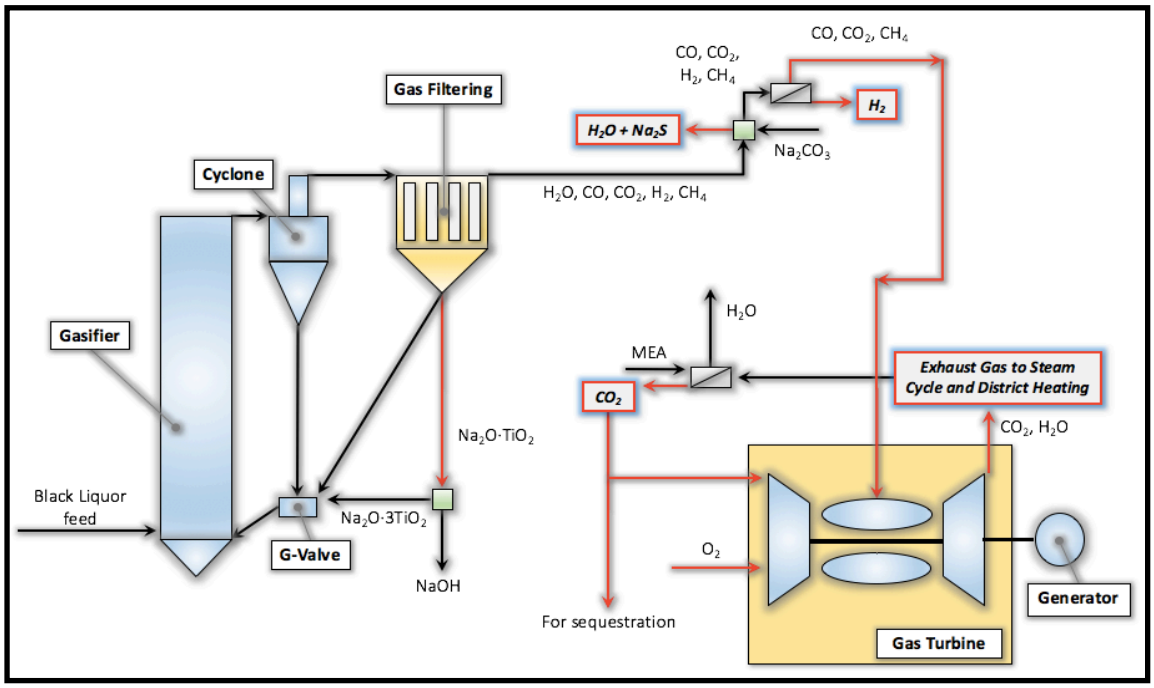


Figure 2. Different solutions can be either integrated as shown in the figure or operated separately

In Table 2, we have summarized the results from the simulations using different conditions with respect to calculated gas composition using the PLS-models. The energy and material balances are performed using the physical model. The physical model takes into account endothermic and exothermic reactions like reduction of SO_4 respective to oxidation of C and H.

The heat transfers through walls, in heat exchanger and scrubber are calculated as well as particle separation in the cyclone and the bag filter. However, the absorption of H_2S respective to CO_2 in the scrubber as a function of pH of the scrubber solution is not included here. The recycling of ash from the bag house is also not included in this simulation.

For the black liquor gasification, we found that there is residual carbon not converted during the gasification. By recirculating the solids from the bag-house into the reactor, the amount of residual carbon dropped to below 4% in the filter ash at stabilized conditions. This resulted in a carbon conversion of higher than 96-99% at steady state with recycled dust from the bag house to the down-comer of cyclone 1.

The results showed that the H_2 content is in the range 9 -17%vol., while the CO content is only 2-4%. The CH_4 content is in the range 1-1.9%, which is quite high in relation to the CO content. Since H_2S is stripped off, the concentration of H_2S is found to be 0.5-0.8% with SO_4 reduction of 87-93%, which is as good or better as in a conventional recovery boiler.

The moisture (H_2O) is calculated from the shift reaction with the constant KT given for the average temperature (T) at steady state conditions as shown in

$$K_T = [CO][H_2O]/[H_2]\,[CO_2] \qquad (1)$$

This is for the actual gasification. For the moisture content in the synthesis gas after the scrubber, the water content of the synthesis gas at saturation for the given scrubber temperature is used. From this we recalculate the gas composition used in the simulations later on as a function of operating conditions, but then combining also with energy and mass balances.

4 System Studies

The experiments are not always easy to control as exactly as wanted in the pilot scale plants. For the analysis of impacts of different variables (like temperature, relative oxidation, organic load and impact of water/moisture for the black liquor), the model still is good enough in relation to other uncertainties. By inserting the conditions and gas composition, the material and energy balances are determined for the systems studied in addition to the dynamics and controllability. The detailed system analysis is not possible to include in this paper due to the limited number of pages available, and thus will be presented in future studies.

For system analysis, one study on H_2 production in a CHP plant is presented in Naqvi et al (2017b). Yang and Ogden (2007) made an overview of production costs for Hydrogen production as well. Another study was made on black liquor gasification systems Dahlquist et al (2017), where different cycles and solutions were compared, including among others CO_2 removal. Asadullah (2014) has made a critical review of downstream gas cleaning after biomass separation, which includes the particle and tar removal.

In Fig. 2, we have a gasification process that could be used for black liquors or another biomass. The BLG

with addition of titanate (TiO_2) can give a solution with direct caustization for conversion of Na_2CO_3 to NaOH, as well as selective absorption of H_2S for the chemical recovery. The gas then can be separated to extract hydrogen, while the residual gas is combusted directly in a boiler, or in an external gas turbine combustor, making a combined cycle possible. Heat from the steam turbine condenser then can be used for e.g. district heating. Even CO_2 can be removed at the far end of the exhaust gas train. The BLG with direct causticization at the pulp mill is an interesting option (Dahlquist, Jones, 2005). The alternative with a combi-cycle could give an electric to fuel heating value efficiency of up to 38%, which is high for a process with such a poor fuel.

5 Conclusions

In this paper, we have presented that how regression models such as PLS, PCA and similar can be made from experiments and combined with dynamic physical models developed in e.g. Modelica. Such developed models can be used to study different systems from the energy and material balance perspective, but also investigate how to go from one process mode to another in a smooth way. This will be more important as the economic conditions will vary much more in the future from one part of the day to another, as well as over the season, making it much more complex to fulfil all different demands. When earlier the focus has been on conversion processes like gasification, we will see an increasing demand also for gas separation like membrane separation for developing efficient system solutions. New demands like CO_2 removal may give quite different economical optima, if CO_2 is valued significantly higher than today. This also will shift the use of fossil fuels for production of chemicals into a demand to use biomass, which will give new incentives to the proposed processes for production of base chemicals like CH_4 and hydrogen.

Acknowledgements

We thank Bioregional and especially Sue Riddlestone, for making their pilot plant available for the tests with wood pellets, and Swedish Energy Agency and KKS are acknowledged for the financial support.

References

A. Gómez-Barea and B. Leckner. Modeling of biomass gasification in fluidized bed. *Progress in Energy and Combustion Science*, 36(4): 444–509, 2010.

C. Di Blasi. Kinetic modeling of biomass gasification and combustion. Intelligent Energy Europe (PyNe). Available via https://pdfs.semanticscholar.org/dddc/d68110b3f918f67a9800a311ab2481db12ef.pdf [accessed January 5, 2016].

C. Roberto and M. D. Veroli. Mathematical Modelling of Biomass Gasification in a Circulating Fluidized Bed CFB Reactor. *Journal of Sustainable Bioenergy Systems*, 2: 160-169, 2012.

C. Yang, and J. Ogden. Determining the lowest-cost hydrogen delivery mode. *International Journal of Hydrogen Energy*, 32: 268-286, 2007.

D. Baruah and D. C. Baruah. Modeling of biomass gasification: A review. *Renewable and Sustainable Energy Reviews*, 39: 806–815, 2014.

E. Dahlquist and A. Jones. Presentation of a dry black liquor gasification process with direct caustization. *TAPPI Journal*: 15-19, 2005.

E. Dahlquist, M. Naqvi, E. Thorin, J. Yan, K. Kyprianidis, and P. Hartwell. Experimental and numerical investigation of pellet and black liquor gasification for polygeneration plant. *Applied Energy*, 204:1055-1064, 2017.

M. Asadullah. Biomass gasification gas cleaning for downstream applications: A comparative critical review. *Renewable and Sustainable Energy Reviews*, 40: 118-132, 2014.

M. Naqvi, E. Dahlquist, and J. Yan. Complementing existing CHP plants using biomass for production of hydrogen and burning the residual gas in a CHP boiler. *Biofuels*, 8(6): 675-683, 2017.

M. Naqvi, J. Yan, M. Danish, U. Farooq, and S. Lu. An experimental study on hydrogen enriched gas with reduced tar formation using pre-treated olivine in dual bed steam gasification of mixed biomass compost. *International Journal of Hydrogen Energy*, 41(25): 10608-10618, 2016.

M. Naqvi, J. Yan, and E. Dahlquist. Synthetic natural gas (SNG) production at pulp mills from a circulating fluidized bed black liquor gasification process with direct causticization. *In Conference proceedings of ECOS 2010.*

M. Naqvi, J. Yan, E. Dahlquist, and S. Naqvi. Off-grid electricity generation using mixed biomass compost: A scenario-based study with sensitivity analysis. *Applied Energy*, 201: 363-370, 2017.

Cascade Optimization using Controlled Random Search Algorithm and CFD Techniques for ORC Application

Ramiro G. Ramirez Camacho[1] Edna R. da Silva [2] Konstantinos G. Kyprianidis[3] Oliver Visconti [4]

[1] Mechanical Engineering Institute, UNIFEI - Federal University of Itajubá,Itajubá, Brazil, ramirez@unifei.edu.br
[2,]Future Energy Center, MDH Mälardalen University, Västerås, Sweden, ednaunifei@yahoo.com.br
[3]Future Energy Center, MDH Mälardalen University, Västerås, Sweden, konstantinos.kyprianidis@mdh.se
[4]Mechanical Engineering Institute,UNIFEI - Federal University of Itajubá, Itajubá, Brazil, oliver@unfei.edu.br

Abstract

This paper presents the methodology for performance optimization of a steam turbine cascade using CFD techniques for ORC (Organic Rankine Cycle) application. The steam turbine cascade is parameterized to achieve the maximum efficiency while using different organic fluids. The main objective of this work is to attain the maximization C_l /C_d ratio from a preliminary design. The approach to finding the maximum C_l /C_d ratio is based in optimization algorithms. The CRSA (Controlled Random Search Algorithm) was chosen for the optimization process. The optimization algorithm (CRSA) is integrated with CFD techniques, using automatic building schemes of parameterized geometries and meshes via "*script files*" with editing commands written in Tlc/Tk language, which will be interpreted by the commercial software ICEM-CFD®, in batch mode. Finally, for the numerical calculation, the commercial software FLUENT® is used with fluid properties, real gas model, turbulence model and boundary conditions set through "journal files". In this paper, R245fa and Toluene are used as working fluids. Results of drag, lift and pressure distribution are reported. This methodology allows making corrections in the initial project of the cascade shape.

Keywords: ORC, optimization, CFD, CRSA, turbomachinery, real gas, equations of state

1 Introduction

The accelerated consumption of fossil fuels has caused many serious environmental problems such as the destruction of the ozone layer, global warming and air pollution. Emissions of carbon dioxide related to energy consumption have increased worldwide from 30.2 billion metric tons to 35.2 billion metric tons in recent years and will be around 43.2 billion metric tons in 2035.

Given that energy resources are becoming more valuable due to the supply and demand relationship and that environmental legislation is becoming stricter, unconventional technologies for energy conversion are necessary to ensure the future supply of electricity. Low-grade heat sources are considered as candidates for new sources of energy.

The technologies for generating electrical power by recovering waste heat sources can be considered a well-established and mature way of energy production; taking into account that the thermodynamic cycles being more exploited are those using steam turbines, with the conventional Rankine Cycle as the most used due to its advantages, such as price, availability and non-toxicity of the working fluid. However, for heat sources with temperatures below 400 0C, it is quite difficult to use water as the working fluid because of the need to apply vacuum in a large part of the plant, making it less efficient and increasing generation costs considerably.

Accordingly, the Organic Rankine Cycle (ORC) is commonly used to generate energy from sources of low and medium temperature (geothermal, waste heat from processes, solar energy, etc.) and in recent decades has extended its use to sources of heat at high temperatures (biomass burning and exhaust gases from primary triggers).

The majority of the researches conducted in this field is focused on extensive analysis of different thermodynamic cycles and working fluids, seeking to develop energy systems more efficient by selecting the most appropriate organic working fluid and an optimal set of operating parameters.

However, the majority of these studies have different insights when defining the optimal set of criteria leading to an optimal configuration of the cycle according to the characteristics of the source of heat available. Furthermore, most of which rule out the possibility of analyzing the effects of the variation of various operating parameters of the cycle have on other indicators of interest, as the size and efficiency of the turbine, for example.

In this sense, researchers have committed a lot of efforts to develop methods for optimal design based on genetic algorithms to find the best design point. Recently, (Lemort et al., 2009) dealt with a method of maximizing the efficiency of the steam turbine based on genetic algorithms (Qin et al., 2003). This method has a number of functions that are taken as constraints. Thus the optimal geometry

and aerodynamic parameters are solved using the genetic algorithm.

Researchers are more and more using CFD techniques because through certain defined geometry and with the use of correct boundary conditions, it is possible to calculate the local and global variables of the flow field.

The fundamental basis of almost every CFD problem is the Navier–Stokes equations, which define any single-phase fluid flow. However, it is not possible to only use CFD techniques when dealing with a great number of geometric and flow parameters. Then, in order to attain the correct solution, it is best to use an optimization algorithm (OA).

2 Organic Rankine Cycle (ORC)

The organic Rankine cycle (ORC) is similar to the conventional Rankine cycle power conversion. However, this system uses a high-density organic compound as the working fluid instead of water.

Saturation curves for water and an organic fluid are presented in Figure 1. The advantages provided by water and the organic fluids in each application are directly related to the observed differences in their saturation curves. More specifically, the large difference between both types of fluid is in the slope of the saturation vapor curves, which directly influences the behavior of the fluid during its expansion through the turbine. In the case of water, the vapor curve displays a negative slope (Figure 1a). However, the vapor curve of many organic fluids, exhibits a positive slope (Figure 1b). The expansion in the turbine takes place for the three types of fluid differently. In the case of water, the saturated vapor enters the turbine and undergoes an isentropic expansion to the condensing pressure of the cycle, the fluid output exhibits a high fraction of liquid (from the point of view of conservation of the internal structure of the turbine). Thus, the employment of overheating and reheat in the cycle becomes indispensable, in order to avoid the deterioration of the equipment, introducing complications to the system design. In the case of an organic fluid, superheated steam is obtained after expansion in the turbine, rather than a liquid-vapor mixture. The absence of liquid along the turbine translates into a simpler system design, since there is no need for superheat and reheats in the cycle. Furthermore, because the fluid exiting the turbine is superheated, its temperature is higher than the condensing temperature, even though its pressure is identical to the condensing pressure. The appearance of the higher temperature, creates a potential heat transfer; enabling the use of part of the existing energy for preheating the fluid in the inlet of the steam generator. By harnessing the energy output of the turbine, it is possible to increase the cycle efficiency.

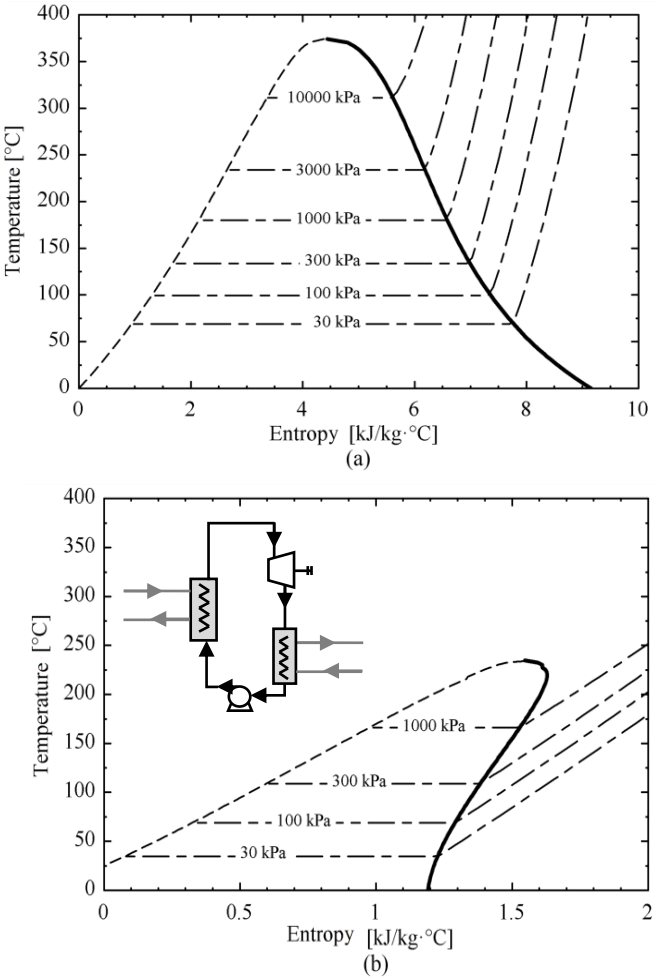


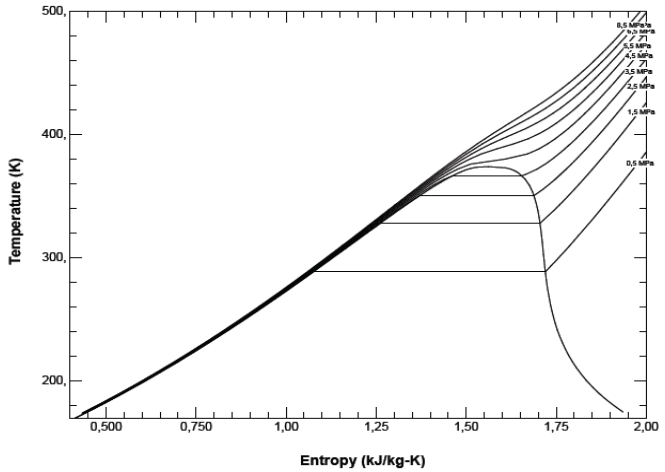
Figure 1. T-s diagrams for water (a) and an organic fluid (b), showing the different inclinations of the saturation vapor curve (continuous line).

In the last two decades, the use and research on this technology have grown rapidly as an option for recovering heat from sources of low temperature and medium temperature, such as solar energy, geothermal energy and waste heat from industrial processes. Today has extended its use for small cogeneration plants using biomass as fuel. Currently, it is expected that applications in industrial processes and modular applications using solar energy will have a rapid development in the coming years. Figure 2 shows one configuration of an ORC cycle, and the processes represented in a T-s diagram for two different flow as R245fa and Toluene.

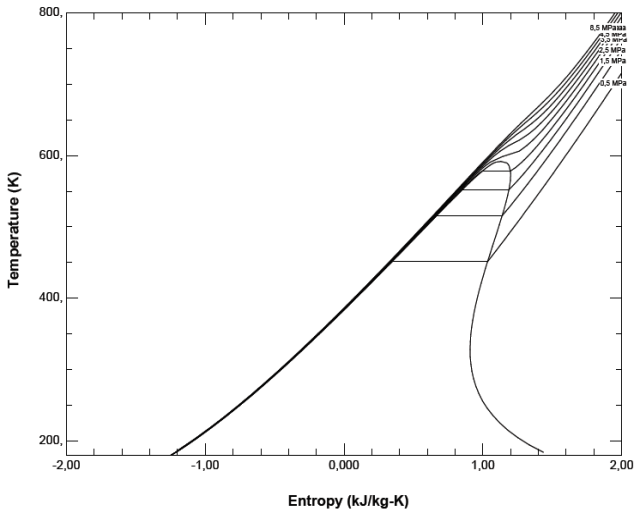
The market facilities ORC is growing apace. Since the installation of the first plants of ORC in the 80s, has been registered an exponential growth in the use of this technology and in Europe currently, there are between 120 and 150 ORC plants (Crowe, 2011). According (Quoilin and Lemort, 2014), the growth in the number of projects and installed power of ORC in the last 20 years had an exponential character of which 48% corresponded to

biomass applications, 31% geothermal, 20% heat sources residual and 1% with solar energy installations.

The ORC system performance is strongly connected to the prime mover, which are classified into two types: positive displacement expanders and turbo machinery. They can be used both axial and radial turbines, being the radial the one that ensure greater isentropic efficiencies in small capacities. The choice of prime mover should be made according to the size of the system. For small applications are used screws and scroll expanders which are in a stage of research and development as seen in the work of (Quoilin et al., 2010; Lemort et al., 2009) and others. While in applications above 200 kW are used axial turbines, with a high degree of technology maturity, which can reach isentropic efficiencies of up to 85%.



(a)



(b)

Figure 2. T–s diagrams showing the different inclinations of the saturation vapor curve to organic fluids, R245fa (a); and to organic fluids, toluene (b).

3 Steam Turbine Design

3.1 Preliminary Turbine Design

The preliminary design of a turbine begins using one-dimensional modeling techniques. The thermo-aerodynamic design of a turbine involves handling a large amount of parameters associated with mechanical calculations to obtain the final geometry for the context in which the turbine is intended. In general, the design consists in the search of some basic geometrical parameters for the rotor blades - the design variables - in order to maximize the turbine efficiency.

In this study, the preliminary design of the turbines was performed using an in-house code and following the design procedure established by (Saravanamutto et al., 2001). The general procedure implemented was to determine the overall dimensions of the machine along with blade and flow angles and isentropic efficiency by fixing the mass flow rate, inlet and outlet pressures as calculated by the cycle analysis and by assuming flow and loading coefficients (Table 1).

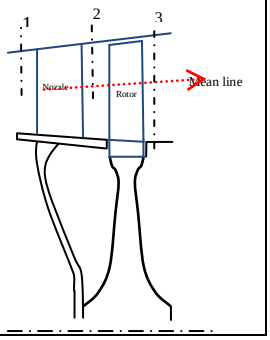
Table 1. Turbine Initial Design Operating Conditions.

Operating Conditions	
Mass flow	m = 20 kg/s
Isentropic efficiency	ηt = 0.9
Inlet temperature	T_{01} = 1100 K
Variation of temperature	ΔT_0=145 = T_{01}-T_{03} K
Variation of Pressure	ΔP=1.873 = P_{01}/P_{03}
Pressure inlet	P_{01} = 4 bar
Rotation	n = 250 rps
Peripheral velocity	U = 340 m/s
Loss coefficient in the stator	λ_N = 0.05
Load Coefficient	ϕ = 0.8

From the basic design specifications and the performance analysis, a refinement of the results was performed in order to optimize the geometry (Table 2).

Table 2. Turbine Final Design Operating and Geometrical Conditions.

	α_1	α_2	α_3	β_1	β_2
Root	0°	62.15°	12.12°	39.32°	51.13°
Mean	0°	58.38°	10°	20.49°	54.96°
Tip	0°	54.93°	8.52°	0°	58.33°
		1	**2**	**3**	**Unit**
P		3.54	2.49	1.856	bar
T_0		1067	982.7	922	k
ρ		1.155	0.883	0.702	kg/m^3
A		0.0626	0.0833	0.1047	m^2
r_m		0.216	0.216	0.216	m
r_t/r_r		1.24	1.33	1.43	---

h	0.046	0.0612	0.077	m
	Stator	**Rotor**		
s/c	0.86	0.83		
h(mean)	0.0536	0.0691 m		
h/c	3	3.00		
c	0.0175	0.0230 m		
s	0.01506	0.0191m		
z_{blade}	90	71		

3.2 Linear Cascade Design

Based on the defined geometry for the rotor, it is possible to generate a linear cascade which represents the axial rotor, considering a line on the average height of the blade, the relative velocity field on the cascade and the associated boundary conditions. Thus, the angles of incidence, stagger angle and pitch cascade are defined in the average height of the blade. Figure 3 illustrates the basic gas turbine cascade configuration.

For the turbine cascade blade design, the camber line was estimated using the directions of relative velocities in the inlet and exit, obtained from the preliminary turbine design. This was done graphically by *script* written in Tcl/Tk language for interpreting by the software ICEM-CFD®. Given the chord length (c) and angles relative (β_1) and (β_2).

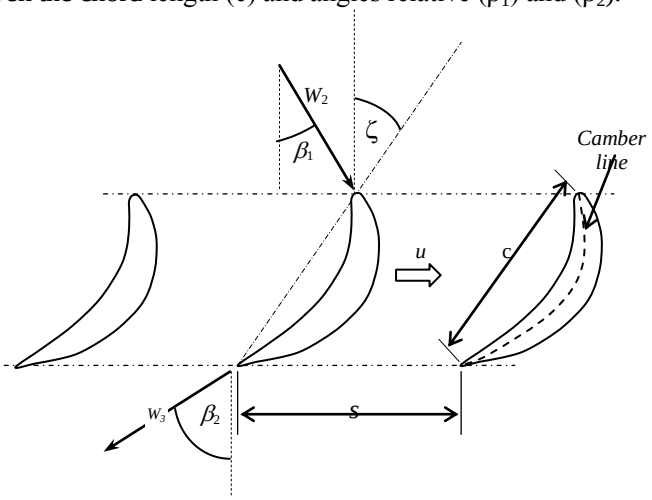


Figure 3. Cascade turbine configuration.

The tangents were brought to an intersection with each other and subdivided into equal distances. The envelope to the inner region of the connecting lines is the camber line. Once the camber line was constructed, a NACA 6519 profile was superimposed and the new profile was generated. The camber line was constructed, a NACA 6519 profile was superimposed and the new profile was generated.

4 Integration Process CFD And Optimization

CFD techniques have been developed over the past decades as a powerful analysis tool for quantification of flow fields in complex geometries, especially those found in Turbomachinery design. Such techniques have been used by the aeronautics industry since the 60ths, begging with the classical panel method with boundary layer interaction to account viscous effects. Nowadays, the use of computational fluid dynamics (CFD) for solving the full Navier-Stokes equations have become a common issue in several industrial design activities. Nevertheless, such computations may represent a bottleneck when a greater number of concurrent geometrical and flow parameters must be analyzed during the searching of good solutions for satisfying certain design objectives. Normally, such task is best accomplished by means of a suitable optimization algorithm (OA). But taking into account real life constraints - the available computational environment and budget - the number of comparative evaluations required by an OA may become prohibitive in a specific design situation (Praveen and Duvigneau, 2007).

To overcome this drawback, several strategies have been conceived for accelerating the optimization task, such as: (i) use of multiprocessing; (ii) use of better optimization algorithms; (iii) use of metamodels (surrogate models) for reducing the number of calls to the true solver model. From a strict engineering point of view, the 3rd strategy seems to be more inexpensive and universal, since it does not rely on costly hardware improvements neither on technical advances in optimization algorithms (da Silva et al, 2012).

4.1 Optimization Process

This work was used as algorithm the optimization stochastic, population-set based algorithm, capable of performing global optimization tasks efficiently, the CRSA, it was first proposed by (Price, 1977) and later improved by (Ali et al., 1997). Further improvements were introduced by (Manzanares et al., 2005).

The CRSA from an initial population of individuals over a consistent region of the problem promotes iterative substitutions of the worst individuals by the best, willing that the population shrink up around the global optimum. The points randomly generated in the space explored, following an iteration process converges to a global minimum by procedures purely heuristic (Ali et al., 1997; Ali and Törn, 2004).

4.2 Flow Calculation in CFD

The blade cascade analysis still represents a fundamental tool in Turbomachinery design context. Relying on 2-D flow models, cascade flow computations are much faster than 3-D models of similar physical complexity. For testing purposes, the CRSA methodology is applied now to a simple case of blade cascade design.

Relevant resulting quantities include the pressure distribution on the blade surface, the flow deflection angle, energy losses, contours of number Mach, lift and drag blade forces.

The flow computations were made using the CFD software FLUENT®. The required meshes were generated by the software ICEM-CFD, through the editing commands in Tcl/Tk.

Prior to optimization process certain steps should be prepared, for example, the definition of a computational domain and mesh generation. The 2-D meshes are generated by a script written in Tcl/Tk language that can be modified by the optimizer and interpreted by the software ICEM-CFD®. Care is taken in the refinement of the mesh near the wall in order to properly quantify the friction stresses. Figure 4 shows the computational periodic zone of linear cascade.

The organic fluid type is defined based on the thermodynamic properties, density and dynamic viscosity. The initial hypotheses, the discretized forms of transport equations are solved iteratively, and the solution must converge.

Real gases, as opposed to a perfect or ideal gas, exhibit properties that cannot be explained entirely using the ideal gas law. The NIST (National Institute of Standards and Technology) real gas model that use the Thermodynamic and Transport Properties of Refrigerants and Refrigerant Mixtures (Lemmon et al., 2006) as an ANSYS FLUENT® shared library (REFPROP v7.0) was used to evaluate the thermodynamic and transport properties of the working fluids.

The REFPROP v7.0 database employs accurate pure-fluid equations of state that are available from NIST. These equations are based on the following three models: Modified Benedict-Webb-Rubin (MBWR) equation of state, Helmholtz-energy equation of state and extended corresponding states (ECS).

In this study will be used to model real gas proposed by Benedict - Rubin. Recently, (Colonna et al., 2006), was analyzed using CDF code, a cascade of a stator of a radial turbine with three different fluid models; the simple polytrophic ideal gas law, the Peng-Robinson-Stryjek Vera cubic EoS and the state-of-the-art Span-Wagner EoS. According (Colonna et al., 2006) the fluid dynamic results are very similar for the computations employing the Span-Wagner and Peng – Robinson Stryjek-Vera EoS. The proposed model MBWR, is very similar to the last.

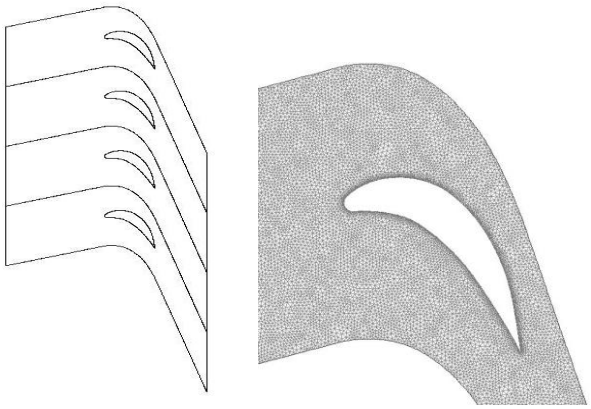


Figure 4. Periodic Channel and Unstructured mesh (40000 cell)

The mass flow or velocity is set at the cascade inlet and the pressure at the cascade outlet. Periodic boundary conditions are considered for reducing the computational domain to a unique periodic region around an airfoil. The turbulence model Spalart-Allmaras (SA) with wall functions is chosen since these enable realistic responses to aerodynamics problems (Azevedo et al., 2003; Spalart and Allmaras, 1992).

The drag and lift coefficients are calculated with basis on the magnitude of the mean velocity vector. The drag coefficient is computed: first, the difference of total pressure between cascade inlet and outlet is evaluated and the following loss coefficient ζr is computed (Vavra, 1974).

$$\zeta_r = \frac{P_1 - P_2}{(\rho/2)W_2^2} \tag{1}$$

The outlet mass average quantities are evaluated by control line (line/rake) located at a distance of a chord length from the trailing edge.

Hence, the drag coefficient is computed by the following relationship:

$$C_d = \frac{\zeta_r \cos^3\beta_\infty}{(s/c)\cos^2\beta_2} \tag{2}$$

This methodology for calculation the drag coefficient avoids numerical errors associated with the integration of the blade surface forces.

The lift coefficient is then computed:

$$C_l = 2(s/c)\left[\tan\beta_1 + \tan\beta_2\right]\cos\beta_\infty - C_d \tan\beta_\infty \tag{3}$$

Several efficiencies are used to compare the performance of the turbine stage; the most common definition is the adiabatic efficiency. However, in cases involving cascade can be used as criteria's of efficiency as total pressure loss

coefficient and diffusion factor. On the other hand, the lift and drag ratio maximum (C_l /C_d), can be used as way evaluating the loading aerodynamic in the cascade.

4.3 Process Integration Methodology

According to (Quoilin and Lemort, 2014), to optimize complex systems, it is necessary to use methods of process integration, this is, CFD flow calculation and optimization algorithms. These methodologies contribute significantly to the development of engineering optimal designs.

The CRSA was adopted as a direct optimizer, an initial population of 10 (n +1) individuals were adopted, where n is the number of design variables (Stagger angle ζ and pitch s). The convergence criterion was 1% (absolute difference between the function values in the worst and best parts of the population or a maximum number of evaluations equal to 500).

The optimization process was obtained by integrating the CRSA with the CFD-Fluent® (CRSA → *script.dat* → ICEM-CFD® → *Journal.file* → Fluent ®). Used, commands in DOS, for running in FORTRAN through the CRSA, a "*script*" of commands in Tcl/tk® is written to interpretation in ICEM-CFD, one computational mesh is generated with variations of stagger angle and pitch for a profile known. Through the script, also provides information about the parameters mesh in regions close to the wall, then a file "*journal.jou*", is edited with information for program execution Fluent: number iteration, criteria convergence, turbulence model, boundary conditions and the formulations for drag and lift coefficients (2) and (3).

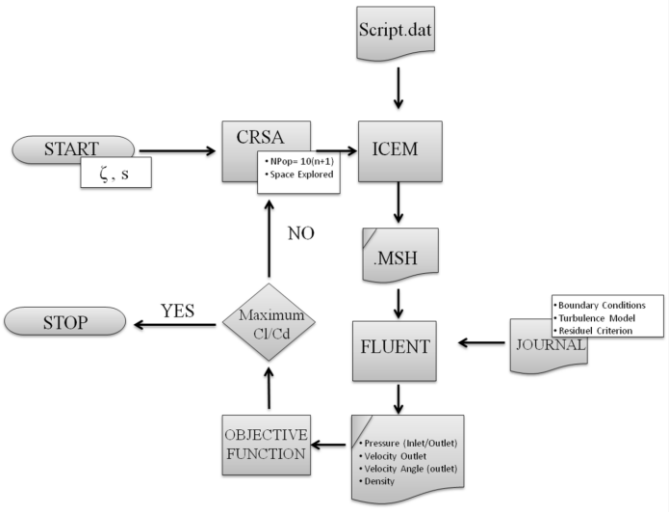


Figure 5. Process Integration Structure.

Have been considered critical temperatures of organic fluids in the boundary conditions at the input as: R245fa (154° C) and Toluene (318.64 °C).

4.4 Optimization Process Results

Table 3 presents the ranges for the generation of the plan of experiments considering the variables of the design: stagger angle and pitch cascade to two organic fluids: R245fa, Toluene.

The optimization process was initiated with a population of 30, NPOP = 10 (n +1), n = 2 (stagger angle and pitch cascade) as criteria convergence was used the residual of ε = 0.0001, value between the two best values found of C_l/C_d in the optimization process. One first analyzes was defined the intervals for stagger angle and pitch cascade.

Table 3. Intervals to the Population Generation

Fluids organic	ζ Stagger angle	s (m) pitch cascade
R245fa and Toluene	41.00 – 44.66	0.0081- 0.015

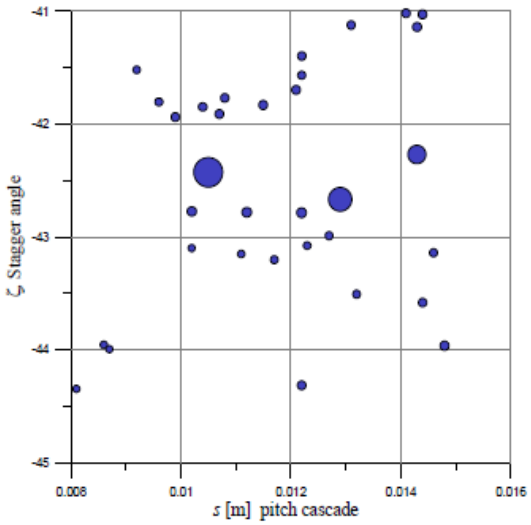


Figure 6. Bubbles plot of C_l/C_d to R245fa.

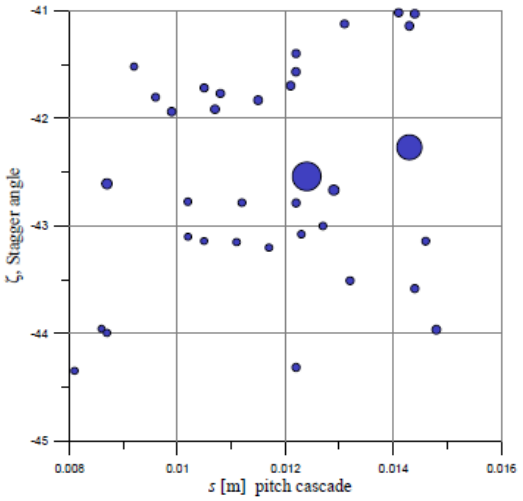


Figure 7. Bubbles plot of C_l/C_d to Toluene.

After the optimization process using the CRSA (Controlled Random Search Algorithm), and considering an initial population of 30 randomly design variable, with the aim of maximizing the relation C_l/C_d, is presented in Figures 6 and 7 in the bubble plot, values C_l/C_d inside the range of variation of stagger angle and pitch cascade for the R245fa and Toluene.

Based on the maximum values of C_l/C_d presented in Figures. 6 and 7, was reduced search interval and also generated a new population with 30 design variables. Table 4 shows the new ranges for the two organic fluids.

Table 4. Intervals for the two Organic Fluids.

	ζ Stagger angle	s (m) pitch cascade
R245fa	42.40- 42.50	0.0081 - 0.015
Toluene	42.00 – 42.60	0.0081 - 0.015

Table 5 present optimum values of the coefficients of drag and lift, quantified by the aerodynamic load, C_l/C_d. Optimization processes are initialized based on the profile NACA 6519, gas turbine cascade.

The results of aerodynamic coefficients are calculated using (2) and (3), where the values of the deflection angles of the flow cascade, and variations in total pressure has been calculated in the inlet and outlet regions of the cascade. Results obtained from the optimization process with CRSA.

Table 5. Results of Optimization Process From CRSA.

Fluid organic	ζ Stagger angle	s (m) pitch cascade	C_l	C_d	C_l/C_d*	C_l/C_d optimum
R245fa	43.2630	0.01025	1.0128	0.0353	28.6931	
	42.4710	0.01440	0.7003	0.0027		262.0300
Toluene	43.2630	0.01025	1.4020	0.0667	21.0283	
	42.4849	0.00860	0.3181	0.0016		198.5575

*Base cascade: NACA 6519; Stagger angle ζ = 43.2630° and pitch, s=0.01025 [m]

In the results presented in Table 5, it is possible to observe that the pitch of the R245fa cascade has greater influence on the increase in the aerodynamic performance. However, the stagger angle, in both fluids, had little influence on the cascade efficiency.

It should be noted also, that the cascade initial design, is based on the cascade calculation of a gas turbine flow (Price, 1977), therefore, a consequence of the values of C_l/C_d ratio has improved considerably from the initial design to the optimized for ORC.

This analyze can be applied to axial rotor, if we consider the medium diameter D_m=0.432 m. then (Z=πD/s), the rotor using R245fa it would have near 94 blades and for Toluene 157 blades. Remember the design the gas turbine was the 71 blades, considering the same relation s/c. This criterion to optimize the load aerodynamic using the C_l/C_d relation is used in cascade plane, but performance in rotor of turbine axial should be applied the isentropic efficiencies in the rotor.

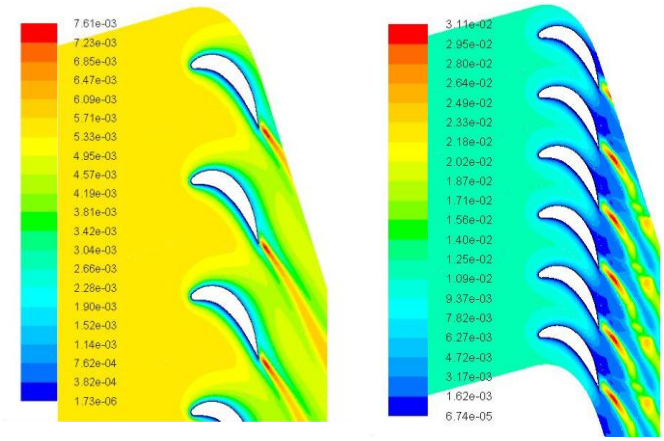


Figure 8. Contours of Viscosity Turbulent, R243fa (a); Contours of Viscosity Turbulent, Toluene(b).

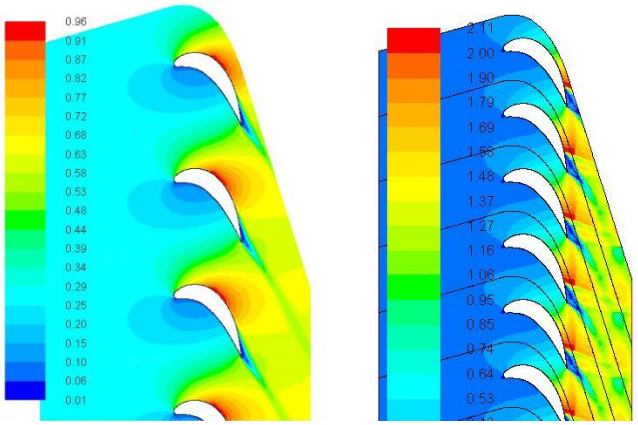


Figure 9. Contours of Mach number, R243fa (a); Contours of Mach number, Toluene (b).

Figures 9a and 9b, shows the contours of viscosity turbulent, where the maximum values are concentrated in the trailing edge region, the toluene cascade present major intensity of vortices. In the Figures 10a and 10b, show the differences in the Mach number, the Toluene cascade can operate with Mach larger than one, frequently this machine operates with high number mach.

5 Conclusions

Based on the design methodology of a gas turbine cascade, it is possible, through optimization techniques and CFD flow calculations, to find an optimum cascade to work with different organic fluids. As a first approach, we analyzed two fluids, with two design variables (pitch and stagger angle). For the optimization process, we used a heuristic algorithm CRSA (Controlled Random Search Algorithm), being effective in finding the optimal solution. Results of a C_l/C_d ratio showing the effects of the design variables (pitch and stagger angle) for two organic fluids were reported. As

can be seen, is necessary optimize the aerodynamic profile changes the configuration of NACA 6519, and introduce the function of camber to correct the separation of the boundary layer in the trailing edge.

Future work will be carried out in order to introduce the tri- dimensional effects on the ORC turbine stator - rotor, aiming to optimize the isentropic efficiencies of this kind of machines. The optimization methodology allows to adapt the preliminary design of gas turbine for the design of cascades with different organic fluids, for analyzing the behavior of the aerodynamic forces and geometric variations in the cascade.

Acknowledgements

This work was carried out with the support of CNPQ/SAAB/CISB and MDH University.

References

M. Ali, A. Törn, and S. Viitanem. A Numerical Comparison of Some Modified Controlled Random Search Algorithm. *Journal of Global Optimization*, 11(4):377-385, 1997. doi:10.1023/A:1008236920512.

M. Ali and A. Törn. Population Set-Based Optimization Algorithms: Some Modifications and Numerical Studies. *Computer and Operations Research*, 31(10):1703-1725, 2004. doi: 10.1016/S0305-0548(03)00116-3.

J. L. F. Azevedo, E. D. V. Bigarella, F. C. Moreira, and E. Basso. Implementação e Teste de Modelos de Turbulência para Aplicações Aeroespaciais em Malhas Não Estruturadas. *In CIBEM 6 - VI Congresso Ibero-Americano de Engenharia Mecânica*, 2003, Coimbra, 1:705-710, 2003.

R. Crowe. Capturing waste heat with rankine cycle systems. *Renewable Energy World.com*, 2011.

P. Colonna, S. Rebay, J. Harinck, and A. Guardone. Real-Gas Effects in ORC Turbine Flow Simulations: Influence of Thermodynamic Models on Flow Fields and Performance Parameters. In *Proceedings of the European Conference on Computational Fluid Dynamics* 2006, Egmond aan Zee, The Netherlands, September, 2006.

E. Lemmon, M. Huber, and M. McLinden. *NIST Standard Reference Database 23: Reference Fluid Thermodynamic and Transport Properties-REFPROP*, Version 9.0, National Institute of Standards and Technology, Standard Reference Data Program, Gaithersburg, 2010.

V. Lemort, S. Quoilin, and C. Pire. Experimental investigation on a hermetic scroll expander. In *Proceedings of the 7th International IIR Conference on Compressors*. 2009.

N. Manzanares-Filho, C. A. A. Moino, and A. B. Jorge. An Improved Controlled Random Search Algorithm for Inverse Airfoil Cascade Design. In *Proceedings of the 6th World Congress of Structural and Multidisciplinary Optimization*, Paper No 4451, Rio de Janeiro, Brazil, 2005.

C. Praveen and R. Duvigneau. *Radial Basis Functions and Kriging Metamodels for Aerodynamic Optimization*. INRIA; 40 pages. RR-6151; France, 2007.

W. L. Price. A controlled random search procedure for global optimisation, *Computer Journal* 20(4):367–370,1977.

X. Qin, L. Chen, F. Sun, and C. Wu. Optimization for a steam turbine stage efficiency using a genetic algorithm. *Applied Thermal Engineering*, 23(18):2307-2316, 2003. doi: 10.1016/S1359-4311(03)00213-8.

S. Quoilin, S. Declaye, and V. Lemort. Expansion Machine and fluid selection for the Organic Rankine Cycle. In *Proceedings 7th International Conference on Heat Transfer, Fluid Mechanics and Thermodynamics*, Antalya, Turkey, 2010.

S. Quoilin and V. Lemort. Technological and Economical Survey of Organic Rankine Cycle Systems. In *Proceedings 5th European Conference Economics and Management of Energy in Industry*, 2014.

H. Saravanamuttoo, G. Rogers, and H. Cohen. *Gas Turbine Theory*, 5th ed. Harlow: Prentice-Hall, 2001.

E. R. da Silva, N. Manzanares-Filho, and R. G. Ramirez Camacho. Metamodelling approach using radial basis functions, stochastic search algorithm and CFD – application to blade cascade design. *International Journal of Mathematical Modelling and Numerical Optimisation*, 3(1/2):87-92, 2012. https://doi.org/10.1504/IJMMNO.2012.044715.

P. Spalart and S. Allmaras. A One-equation Turbulence Model for Aerodynamics Flows. *AIAA Aerospace Sciences Meeting & Exhibit, 30., Reno.* In *Proceedings Washington*, DC: AIAA, 1992.

M. Vavra. *Aero-Thermodynamics and Flow in Turbomachines*, Reprint of the ed. Published by Wiley, New York, ISBN 0-88275-189-1.

Simulation of Light Oil Production from Heterogeneous Reservoirs

Well Completion with Inflow Control Devices

Arash Abbasi , Britt M. E. Moldestad
Department of Process, Energy and Environmental Technology, University of Southeast Norway, Porsgrunn, Norway.
Britt.Moldestad@usn.no arashabbasi84@gmail.com

Abstract

Water breakthrough is a big challenge in light oil production, and different types of inflow control devices are developed to delay or reduce breakthrough. Light oil production from a heterogeneous reservoir is simulated to study the effect of three types of inflow control devices, one passive control and two autonomous controls. NETool is used as the near-well simulation tool. The functionality of passive inflow control device (ICD) and the autonomous rate control production device (RCP) is included in NETool, whereas the autonomous inflow control valve (AICV) is simulated based on expected behaviour. The total production rates and the water cut versus drawdown and the performance curves for ICD, RCP and AICV are studied. The results confirm that RCP and AICV reduce the water production and water cut significantly. The water cut is about 27% for RCP and 44% for ICD at 15 bar. AICV is designed to close 99% for water, and produces negligible amounts of water. The RCP completed well produces about 310 m^3 oil and 110 m^3 water per day at drawdown 15 bar. ICD produces about 230 m^3 water per day, whereas AICV produces insignificant amount of water. The results confirm that the water production decreases with RCP and AICV compared to ICD. Delayed and reduced water production will result in increased oil recovery.

Keywords: light oil, heterogeneous reservoir, ICD, RCP, AICV, water cut, water breakthrough

1 Introduction

A major challenge in oil production is to increase the ability to recover the residual oil. Estimates show that although the oil is localized and mobile, more than half of the oil is remaining in the reservoir after shut down. There are different challenges regarding increasing the oil recovery, and the biggest challenge is water and gas breakthrough to the well. In this study, only water breakthrough in a heterogeneous light oil reservoir will be considered. Oils are categorized based on the density or °API (American Petroleum Institute) of the oil. API gravity is calculated by using the specific gravity of oil, which is defined as the ratio of oil density to the density of water. Light oil is specified by low viscosity, low specific gravity and high °API gravity.

New technology can increase the recovery in new and mature fields significantly. Production data from Statoil's horizontal pilot wells on the Norwegian Continental Shelf show that the cumulative oil production increased with about 20% when new inflow control technology was implemented (Halvorsen *et al.*., 2012) Reservoir and near-well models to show the potential of implementing the new technology are important in order to speed up the implementation of new completion technology. The near-well simulation tool NETool, is used in this study.

1.1 Well completion

Different types of passive Inflow Control Devices (ICDs) have been installed in a number of oil fields all over the world and the implementation has contributed to increase the oil production and recovery significantly compared to open-hole wells (Al-Khelaiwi 2007; Krinis *et al.*, 2009). Newer technology called Autonomous Inflow Control devices (AICDs) has the potential to increase the oil production and recovery even more. Halliburton, Statoil and InflowControl AS have developed AICDs based on different principles (Least *et al.*, 2012; Mathiesen *et al.*, 2011; Aakre *et al.*, 2013). Near well simulations with AICV completion, show high potential regarding increased oil recovery (Aakre *et al.*, 2013). Statoil has currently several wells completed with RCP at the Troll field where the purpose of the RCP is to restrict the gas and maintain the oil production. Results show that the cumulative oil production with RCP completion is 20% higher than a corresponding branch completed with ICDs. Reservoir simulations carried out prior to the installation of RCP, indicated an increased oil production up to 15% (Halvorsen *et al.*, 2012).

1.2 Water breakthrough

Early water breakthrough and high water production result in early shut down of oil wells and low oil recovery. Long horizontal wells are used to obtain maximum reservoir contact. Due to frictional pressure drop along the long well, the driving forces are different from one location to another. This is called the heel to toe effect. In a homogeneous reservoir, the oil production rate can be significantly higher in the heel than in the toe, and this may lead to early water or gas breakthrough in the heel. The heel toe effect is

demonstrated in Figure 1. In heterogeneous or fractured reservoirs, early breakthrough will occur in the high permeable zones due to low flow resistance in the reservoir. Figure 2 presents a fractured reservoir, where the water has high mobility and flows easily to the production well. This is also the challenge in high permeable zones in heterogeneous reservoirs.

New and improved inflow control technology that can spontaneously choke or close for unwanted fluids can solve the problem with early gas or water breakthrough.

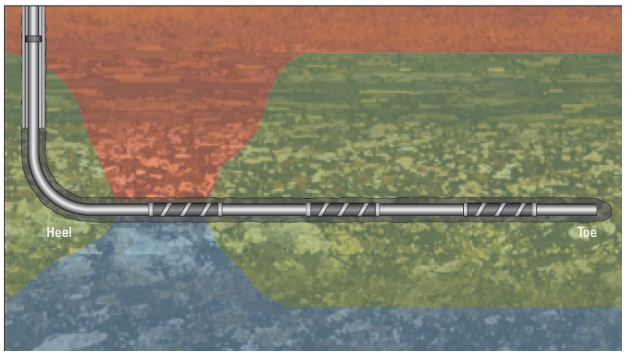


Figure 1. Heel toe effect. Water (blue) and gas (red) breakthrough in the heel (Ellis *et al.*, 2010)

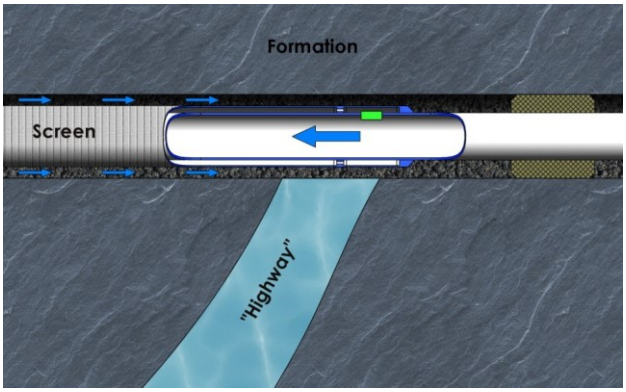


Figure 2. Fractured reservoir with water breakthrough (Aakre *et al.*, 2014)

2 Inflow control devices

Different types of inflow control devices are developed. One passive and two autonomous inflow control devices are described in this chapter.

2.1 Passive inflow control devices

Different types of passive Inflow Control Devices (ICDs) are developed to delay the early breakthrough by restricting the flow. In this study oil production, using nozzle ICD is studied. Well completion with ICDs includes a large number of ICDs evenly distributed along the well. ICDs are designed to give a more uniform oil production along the well. The diameter of each nozzle is chosen to obtain the desired pressure drop over the ICD at a specific flow rate. The pressure drop highly depends on the nozzle diameter and the density of the fluid and less on the viscosity. Passive ICDs are capable of delaying the water breakthrough significantly (Al-Khelaiwi, 2007), and the technology has opened up for production from reservoirs with thin oil columns. The total oil recovery increases significantly with use of ICDs. However, ICDs neither choke nor close for the undesired fluids like water, and after breakthrough, the whole well has to be choked to avoid the downstream separation facilities to be overloaded. Reservoir simulations have been performed for different types of ICD completions and the results have been useful to improve the ICD design (Krinis *et al.*, 2009). Krinis *et al.* used the reservoir model NETool to determine the optimal number and location of ICDs along the well, and they stated that the simulations were the key factor to succeed in optimization of the horizontal well performance. The principle behind the nozzle ICD is based on the following equations (Al-Khelaiwi, 2007):

$$\Delta P = \frac{\rho v^3}{2C^2} = \frac{\rho Q^2}{2A_{valve}^2 C^2} = \frac{8\rho Q^2}{\pi^2 D_{valve}^4 C^2} \quad (1)$$

$$C = \frac{C_D}{\sqrt{(1-\beta^4)}} = \frac{1}{\sqrt{K}} \quad (2)$$

$$\beta = \frac{D_2}{D_1} \quad (3)$$

where ΔP is pressure drop across orifice, ρ is the average fluid density, v is the fluid velocity through an orifice, Q is the fluid flow rate through the orifice, A is the cross section area of orifice, D is the diameter of the orifice, C is the flow coefficient, C_D is the discharge coefficient and K is the pressure drop coefficient.

2.2 Autonomous inflow control devices

In addition to the heel-toe effect that is initializing the coning, coning also occurs due to heterogeneities in the reservoir. Robust inflow control that can choke back and/or close locally the water producing zones has the potential to increase the oil recovery significantly compared to standard ICDs. Statoil has installed one type of Autonomous Inflow Control Devices (AICDs) called Rate Controlled Production (RCP) in wells in the Troll field (Halvorsen *et al.*, 2012). The RCPs delay gas and water coning and in addition, the RCPs have the capability to choke for low viscous fluids after breakthrough. Halliburton has developed an AICD that behave like a traditionally ICD before breakthrough, and choke for the low viscous fluids after breakthrough (Least *et al.*, 2012). The autonomous function of the AICDs, enables the wells to produce for a longer period of time, and the total oil production and oil recovery from a given field will increase (Least *et al.*, 2012). The AICDs are installed in the wells in the same way as the ICDs, and are suitable for production in long horizontal wells. In this study, simulations have been performed with Statoil's RCP.

The RCP is characterized by being very little sensitive to changes in differential pressure, and gives a more uniform flow rate over a range of drawdowns compared to the ICD. The following equations describe the functionality of the RCP (Halvorsen *et al.*, 2012; Mathiesen *et al.*, 2011):

$$\delta P = f(\rho,\mu) \,.\, a_{AICD} \,.\, q^x \tag{4}$$

$$f(\rho,\mu) = \left(\frac{\rho_{mix}^2}{\rho_{cal}}\right) \cdot \left(\frac{\mu_{cal}}{\mu_{mix}}\right)^y \tag{5}$$

$$\rho_{mix} = \alpha_{oil}\,\rho_{oil} + \alpha_{water}\rho_{water} + \alpha_{gas}\rho_{gas} \tag{6}$$

$$\mu_{mix} = \alpha_{oil}\mu_{oil} + \alpha_{water}\mu_{water} + \alpha_{oil}\mu_{oil} \tag{7}$$

where δP is pressure drop through RCP, q is the flow rate, x and y are user input constants, a_{AICD} is the valve strength parameter, α is the volume fraction of the actual phase, ρ_{cal} and μ_{cal} are the calibration density and viscosity.

InflowControl AS has developed an autonomous inflow control valve (AICV) which is completely self-regulating and does not require any electronics or connection to the surface. AICV gives low flow restriction for oil production and has the ability to close almost completely for water and gas. The valves will locally close in the zones with gas and/or water breakthrough, and simultaneously produce oil from the other zones along the well. The AICV technology utilize the fact that flow behavior through laminar and turbulent flow elements are different. The AICV technology consists of two different flow restrictors placed in series. The first one is a laminar flow restrictor and the second is a turbulent flow restrictor. Figure 3 presents a sketch of the combination of flow restrictors, where 1 is the laminar flow element and 2 is the turbulent flow element. The pressure in chamber B activates the piston in the valve to close or open. If oil is flowing through the AICV, the pressure drop through the laminar flow element is high, resulting in low pressure in chamber B and the valve is open. Water gives lower pressure drop through the laminar restrictor, resulting in high pressure in B, and the valve closes.

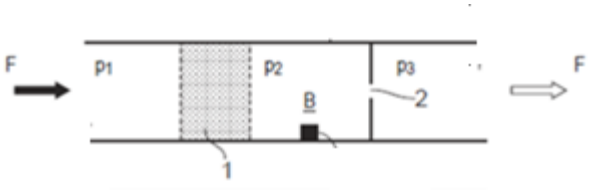


Figure 3. Combination of laminar and turbulent flow restrictors in series

The pressure drops through laminar and turbulent flow elements are expressed by Eq. (9) and (10) respectively. (Aakre *et al.*, 2013; Aakre *et al.*, 2014) The laminar flow element is considered as a pipe segment, and the pressure drop through the element is expressed as:

$$\Delta P = f \,.\, \frac{L.\rho.v^2}{2D} = \frac{64}{Re} \,.\, \frac{L.\rho.v^2}{2D} = \frac{32.\mu.\rho.v.L}{D^2} \tag{8}$$

where ΔP is the pressure drop, f is the laminar friction coefficient, ρ is the fluid density, μ is the fluid viscosity, L is the length of the laminar flow element, D is the diameter of the laminar flow element, Re is Reynolds number.

The pressure drop through the turbulent flow element is proportional to the density and the velocity squared, and is given as:

$$\Delta P = k.\frac{1}{2} \,.\, \rho \,.\, v^2 \tag{9}$$

where k is a geometrical constant, ρ and v is the fluid velocity.

AICV is a new technology and is still not included as an option in NETool. However, AICV has the same function as ICD in open position, and when closed, the flow rate through the valve is reduced to less than 1%. This relationship between open and closed valve is used to simulate the AICV functionality.

Near-well simulations are important to be able to anticipate or predict the economic potential of well completion with different types of inflow controllers. The near-well simulation tool NETool is used in this project.

3 NETool

NETool is a one dimensional steady state near-well simulation tool. The NETool models include fluid properties, reservoir properties and well completion. The required information is imported or defined by the user via a graphical user interface. NETool evaluates the logic and algorithms.

Regarding completion, different options are included in NETool. In this study, different types of inflow controllers are installed in a long horizontal well. The design parameters are specified by the user to fit the specific reservoir and fluid conditions. The well is divided into zones, and the user specifies reservoir and fluid properties for each zone. In addition, the user specify the implementation of inflow controllers, packers, etc. for each zone. The most important user defined inputs to NETool are described below.

3.1 Relative permeability

In numerical reservoir simulation, the relative permeability is significant to predict the oil, water and gas production during the reservoir operation. It is a big challenge to estimate the relative permeability curves for a given field. Relative permeability curves are determined based on experimental core plug tests, and models for relative permeability are developed based on the experimental data.

The relative permeability, K_i, is defined as the effective permeability divided by the absolute reservoir

permeability. Darcy's law describes the absolute reservoir permeability as:

$$q = -\frac{k \cdot A}{\mu} \cdot \frac{dp}{dL} \quad (10)$$

where q is the volume flow rate, k is the absolute permeability, dp/dL is the pressure gradient, A is the reservoir cross section area and μ is the fluid viscosity.

The relative permeability is the ratio between the effective permeability and the absolute permeability, and is a function of the saturation of the different phases in the reservoir. (Selley, 1998; Ahmed, 2006)

$$k_{r,i} = \frac{K_i}{k} \quad (11)$$

where $k_{r,i}$ is the relative permeability for phase i.

In this study, the Corey model and the Stone II model are used to define the relative permeability curves for water and oil respectively. The Corey model [10] for predicting the relative permeability of water is given by:

$$k_{rw} = k_{rowc}\left(\frac{S_w - S_{wc}}{1 - S_{or} - S_{wc}}\right)^{n_w} \quad (12)$$

where k_{rw} is the relative permeability of water, k_{rowc} is the relative permeability of water at the maximum water saturation, S_w is the water saturation, S_{wc} is the irreducible water saturation, S_{or} is the residual oil saturation and n_w is the Corey fitting parameter for water.

The Stone II model estimates the relative permeability of oil in an oil-water system based on the following equation [11]:

$$k_{row} = k_{rowc}\left(\frac{S_w + S_{or} - 1}{S_{wc} + S_{or} - 1}\right)^{n_{ow}} \quad (13)$$

where k_{row} is the relative oil permeability for the water-oil system, k_{rowc} is the endpoint relative permeability for oil in water at irreducible water saturation and n_{ow} is a fitting parameter for oil. (Li &Horne, 2006)

3.2 Input to NETool

Different types of passive and autonomous inflow control devices are available in NETool. In this study the nozzle ICD and Statoil's RCP are utilized. The functionality and equations for these devices are presented in Chapter II. The diameter of nozzle ICD is set as 4 mm. The design parameters for the RCP, x, y and a_{AICD} are set as 4, 1.1 and 10^{-7} respectively.

AICV is a new technology and is still not given as an option in NETool. However, AICV has the same function as ICD in open position, and when closed, the flow rate through the valve is reduced to about 1% of the flow rate in open position. This relationship between open and closed is used to simulate the AICV functionality in NETool.

A sketch of the base-pipe including annulus, inflow control devices and packers are presented in Figure 4. The packers are installed to isolate the different zones, and thereby avoid annulus flow from one zone to another. The well has a total length of 500 m, and is divided 32 zones, with two inflow-controllers in each zones. Each section isolated with packers, includes three or two inflow zones. Three different cases are simulated, one with nozzle ICD, one with RCP and one with AICV.

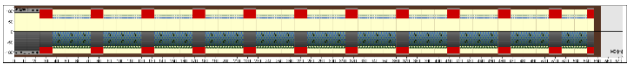


Figure 4. Well completion including packers (red squares) and inflow control devices (black dots).

Figure 5 represents the reservoir permeability along the production well. The reservoir is heterogeneous and has two high permeability zones with permeability 1D, and the permeability in the other is 100 mD. Figure 6 shows the oil saturation in the reservoir. The oil saturation is assumed as 100% in the low permeability zones, whereas the water saturation is assumed 100% in the high permeability zones. Since NETool is a steady state simulation tool, it is not possible to study the changes in oil and water saturation with time nor is it possible to determine the breakthrough time. These simulations are therefore based on the assumption that the water breakthrough has already occurred in the high permeable zones, whereas the low permeable zones are still saturated with oil. This simplification of the oil and water saturation in the reservoir is made to be able to study the effect of the different inflow control devices after water breakthrough.

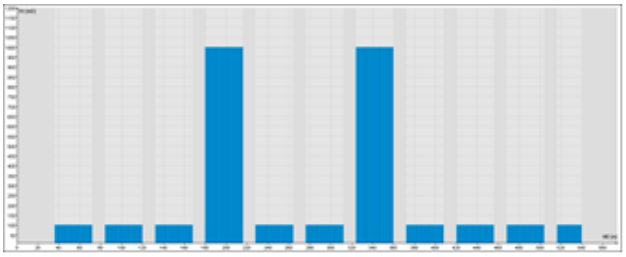


Figure 5. Permeability. The permeability is 1D and 100 mD in the high and low permeability zones, respectively.

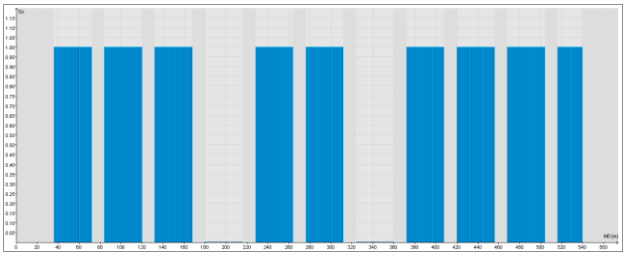


Figure 6. Oil saturation. The low permeability zones are saturated with oil (100%) and the low permeability zones are saturated with water (0% oil).

Tab. 1 represents a summary of the input parameters used in this study, and the estimated relative permeability curves are presented in Figure 7.

Table 1. Input to NETool

Reservoir Parameters and well specifications	
Well length	550 m
Reservoir thickness	200 m
Reservoir width	4000 m
Reservoir Pressure	302 bar
Porosity	0.20
Permeability	100 mD and 1D
Oil viscosity	2cP
°API gravity	33
Reservoir Temperature	68°C
Dissolved gas/oil ratio	**130^3/Sm^3**

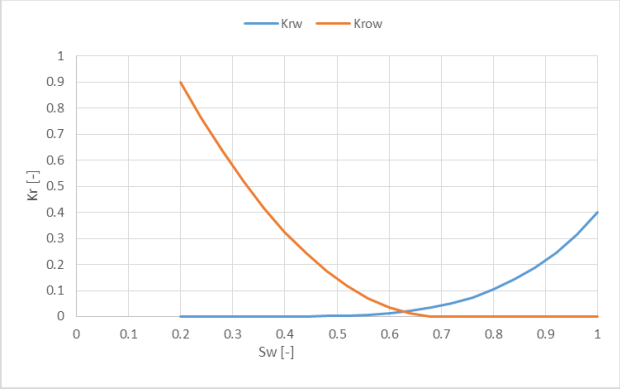


Figure 7. Relative permeability curves for oil-wetted reservoir.

4 Results and discussion

The aim of this study is to find the effect of different types of inflow control devices on oil and water production. Three different cases are performed, one with ICD completion, one with RCP completion and one with AICV completion. The input parameters for the simulations are the same for the three cases. The simulations are run with drawdown of 2, 5, 10 and 15 bar, and the total production rates versus drawdown, the water cut versus drawdown and the performance curves for ICD, RCP and AICV are studied.

These shortcuts Figure 8 shows the oil performance curves for ICD, RCP and AICV. The curves are calculated based on the total production rate from the well and presented as the volume flow of oil through one inflow control as a function of pressure drop over the completion. Due to low permeability in the oil zones, the pressure drop in the reservoir is high which gives a low differential pressure, ΔP, over the inflow controls and low production rates. At the actual ΔP over the inflow control devices, RCP gives the lowest pressure drop per volume flow. The oil flow rate through RCP is about 500 l/h at differential pressure 0.7 bar. ICD and AICV produces less than 500 l/h at ΔP 1.15 bar and 1.25 bar respectively. Figure 9 shows the typical functionality of ICD and RCP. The curves are plotted based on the equations presented in Section 2, and shows that RCP has higher production rates versus ΔP at low ΔP, whereas above a certain pressure it changes. The reason is that ΔP is proportional to the volume flow in the power of 4 for RCP and proportional to the volume flow squared for ICD. Figure 10 gives the comparison between the water performance curves for ICD, RCP and AICV. The water is produced from the high permeability zones where the flow restriction through the reservoir is low. This results in high production rates of water and high pressure drop across the completion. The figure shows that the RCP is choking for water and that the ICD is producing significantly higher amounts of water compared to RCP at ΔP above 1 bar. The functionality of AICV is to close almost completely for water, and at closed position, the flow rate will be about 1% of the flow rate through open valve. Since AICV is not included in NETool, the AICV was simulated by using 4 mm ICDs in the oil zones and 0.4 mm ICDs in the water zones. This will indicate the potential of AICVs. The water curve for AICV shows that the amount of water flowing through the AICV is insignificant.

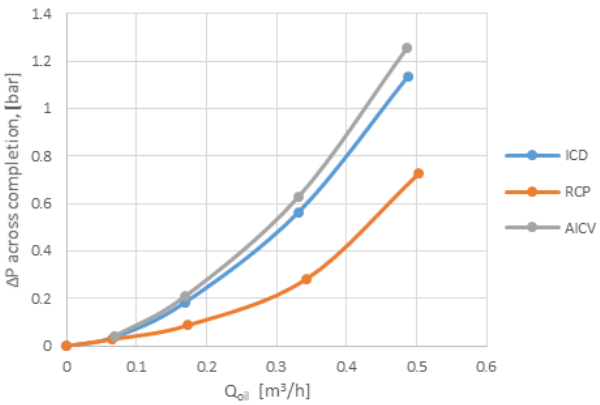


Figure 8. Oil production as a function of differensial pressure through the inflow controllers.

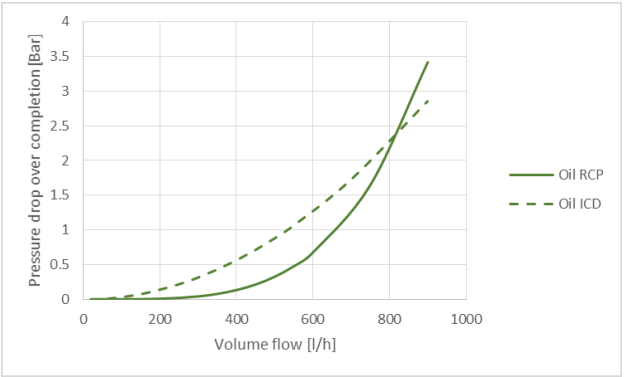


Figure 9. Calculated performance curves for RCP and ICD.

Figure 11 presents the water cut versus drawdown when using ICD, RCP and AICV. The water cut

decreases with increasing drawdown, and at drawdown 15 bar the water cut is about 27% for RCP and 44% for ICD. The water cut differs less at low drawdowns, and at 2 bar, the water cuts are 59% and 64% for RCP and ICD respectively. When the AICV technology is used, the water cut is negligible for the range of simulated drawdowns.

The total production rates for the horizontal well are presented in Figure 12. The figure shows that ICD is producing more water than oil when the drawdown is lower than 10 bar. At higher pressure drops, the oil production is higher than the water production. However, the well is producing water from only 6 zones and oil from 26. RCP is designed to choke for water, and the oil production exceeds the water production at drawdown higher than 3 bars. The ratio between the oil and water production depends on the relative permeability, the fluid properties and the fuctionallity of the inflowcontrol devices. Since ICD is a passive inflow control, the water flow will not be restricted, and water through each ICD will be produced at higher flow rates than oil. RCP is autonomous and chokes for water, which results in low water production and unrestricted oil production at the range of drawdowns used in this study. The simulations are able to predict the benefits of using autonomous inflow control devices.

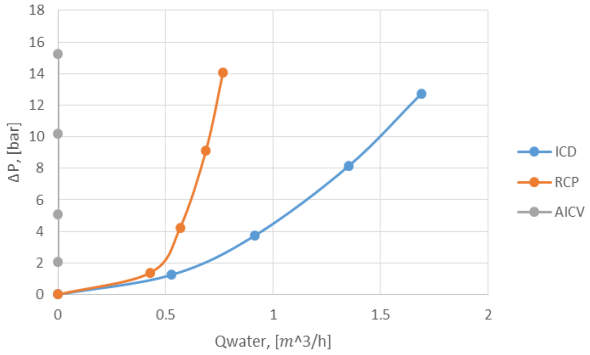


Figure 10. Water production through ICD, RCP and AICV

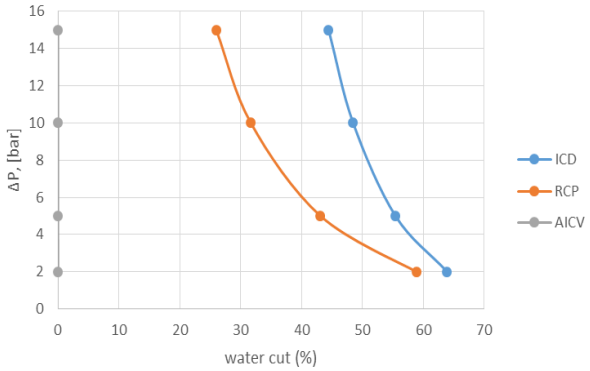


Figure 11. Watercut as a function of drawdown.

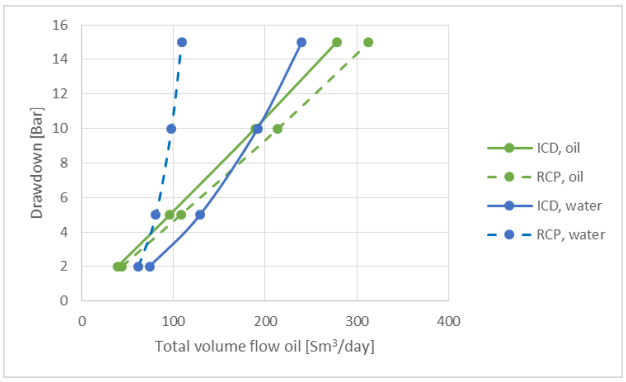


Figure 12. Total production rate of oil and water as a function of drawdown.

5 Conclusions

Water breakthrough is a big challenge in light oil production, and different types of inflow control devices are developed to delay, reduce or avoid breakthrough. Light oil production from a heterogeneous reservoir is simulated to study the effect of the three types of inflow control devices, nozzle ICD, autonomous RCP and AICV. NETool is used as the near-well simulation tool. The functionality of ICD and RCP is included in NETool, whereas AICV is simulated based on expected behaviour. The simulated horizontal well is 550 m long and packers and inflow control devices were evenly distributed along the well. The wells with ICD, RCP and AICV completion were simulated using different drawdowns ranging from 2 to 15 bar. The total production rates versus drawdown, the water cut versus drawdown and the performance curves for ICD, RCP and AICV were studied. The results confirm that autonomous inflow controls, RCP and AICV, reduce the water production and water cut significantly compared to passive ICD. The water cut decreases with drawdown, and is about 27% for RCP and 44% for ICD at 15 bar. When the AICV technology is used, the water cut is negligible for the range of simulated drawdowns. RCP gives the highest oil production rate at drawdown ranging from 2 to15 bar, but this is expected to change when the drawdown is further increased. The RCP completed well produces about 310 m^3 oil and 110 m^3 water per day at drawdown 15 bar. ICD produces about 230 m^3 water per day, whereas AICV produces a negligible amount of water. The results confirm that the water production decreases with RCP and AICV compared to ICD. Delayed and reduced water production will result in increased oil recovery.

References

H. Aakre, B. Halvorsen, B. Werswick, and V. Mathiesen. Autonomous Inflow Control Valve for Heavy and Extra-Heavy Oil. In *SPE 171141, SPE Heavy and Extra Heavy Oil Conference - Latin America, Medellin, Colombia,* September, 2014.

H. Aakre, B. Halvorsen, B. Werswick, and V. Mathiesen. Smart well with autonomous inflow control valve technology. In *SPE 164348-MS, SPE Middel East Oil and Gas Show and Exhibition, Manama, Barhain,* March, 2013.

Tarek Ahmed, *Reservoir Engineering Handbook.* Elsevier GPP, 3. Ed. 2006.

F.T. Al-Khelaiwi and D.R. Davies. Inflow Control Devices: Application and Value Quantification of a Developing Technology. In *International Oil Conference and Exhibition, Veracruz, Mexico,* June 2007.

T. Ellis, A. Erkal, G. Goh, T. Jokela, S. Kvernstuen, and E. Leung. Inflow Control Devices- Raising Profiles. *Oilfield Review*, 21(4):30-31, 2010.

M. Halvorsen, O. M. Nævdal, and G. Elseth. Increased oil production by autonomous inflow control with RCP valves. In *SPE 159634, SPE Annual Technical Conference and Exhibition. San Antonio, Texas, USA,* October, 2012.

D. Krinis, D. Hembling, N. Al-Dawood, S. Al-Qatari, S. Simonian, and G. Salerno. Optimizing Horizontal Well Performance in Non-Uniform Pressure Environments Using Passive Inflow Control. In *OTC 20129, Houston, Texas,* May 2009.

B. Least, S. Greci, R. Burkey, A. Utford, and A. Wileman. Autonomous ICD Single Phase Testing. In *SPE 10165, SPE Annual Technical Conference and Exhibition, San Antonio, Texas, USA,* October 2012.

K. Li and R. N. Horne. Comparison of methods to calculate relative permeability from capillary pressure in consolidated water-wet porous media. *Water resources research,* 42: pp. 285-293, 2006.
doi: 10.1029/2005WR004482.

V. Mathiesen, H. Aakre, B. Werswick, and G. Elseth. The Autonomous RCP Valve-New Technology for Inflow Control in Horizontal Wells. In *SPE Annual Technical Conference and Exhibition, Aberdeen, UK,* 2012.

Richard G. Selley. *Elements of Petroleum Geology*, Academic Press, 2nd edition, 1998.

Functionality Testing of Water Pressure and Flow Calculation for Dynamic Power Plant Modelling

Timo Yli-Fossi
Department of Automation Science and Engineering, Tampere University of Technology, Finland

Abstract

Water pressure and flow rate calculation in dynamic boiler models is challenging because of stiff system dynamics meaning that time constants of model states vary by several orders of magnitude. Furthermore, strong interconnections between pressures and flow variables may cause instability problems in simulation runs. This study presents a method to implement and test dynamic thermal power plant water-steam system models. A dynamic water-steam system model is presented. The model is applied for testing of the functionality of the presented computation model. Computational performance was tested using different numerical solvers. Also sensitivity to changes in initial values of system states and model parameters was tested. The results indicate that a workable way to make flexible models was found.

Keywords: modelling, simulation, power plant

1 Introduction

Water is the working fluid of steam power plants. Different physical phases of water set challenges to modelling work. Process components can contain sub-saturated liquid, saturated liquid-vapour mixture and superheated steam. In supercritical boilers supercritical fluid is also present. Physical characteristics of water, such as density and vapour fraction, change as functions of fluid pressure and enthalpy. Water properties affect also to pressure losses and heat transfer coefficients between the fluid and the internal surfaces of the process components. Temperature differences and flow rates affect heat flows. Likewise pressures and flows are dependent on each other. Therefore, the strong interconnections between the variables, nonlinearity and time variant phenomena set challenges to the modelling and simulation work.

Mathematically, dynamic pressure and flow models are stiff systems of differential equations. Generally, stiffness can be described as a property that existing time constants of the system states differs from each other by several orders of magnitude. A basic difficulty in numerical solutions of stiff models is a requirement of absolute stability. Numerical problems may result to noise in simulation results, which further may lead simulation run to fail. (Åman, 2011)

Besides the stiffness, computational problems are caused by calculation errors and discontinuities. Truncation and rounding errors are caused by digital computing. The truncation error forms a difference between the numerical and real solution. The rounding error results from limited calculation accuracy. Extension of the simulation step length will expand the truncation error but reduce the rounding error when the amount of calculation operations is reduced. If the time step has been selected too long, the solution may begin to oscillate and calculated values to drift away from an allowed area. (Jäntti, 1996)

The differential and/or partial differential equations of the dynamic models can be non-linear, contain discontinuities and consist of complex boundary conditions. Hence, numerical solvers must be applied in simulations. The choice of the solver has significance for simulation speed and accuracy. Numerical methods are often classified in two groups: explicit and implicit. Explicit methods calculate the next state of a model from the current state. Implicit methods find a solution by solving equations involving both the current state and the next one. Implicit methods are usually more efficient when solving stiff models. Implicit methods require an extra computation, but on the other hand the time step to be used can often be lengthened so that the stability of the solution does not suffer. (Åman, 2011; Jäntti, 1996)

Solvers can also be divided into two main types: fixed-step and variable-step solvers. The solvers, which use the variable step, adjust the step length according to the situation. During a fast transient situation, a minor step size is required in the integration. When the model has slow dynamics, near steady state situation and the changes are small, the use of the longer time step is reasonable.

Special situations such as small flows or quick pressure changes may also cause computational problems. The pressure and flow models may also be sensitive to initial values of state and model parameters. This complicates the revision of the models.

Instabilities in modelled pressure and flow, which are not caused by computational reasons, may also be found. This brings the extra challenge for the examination of the instabilities of the model. For example, oscillations can cause problems in control system, boiling crisis and excursion of flow due to differences in the pressure drop characteristics of different flow pattern. Pressure and flow instabilities exist especially during startups and shutdowns. (Majuri, 2012)

The objective of this study was to find the workable way to implement reliable and stable models. The models must be also flexible and easily edited. The water pressure and flow model has been developed for functionality testing. Additionally, the sensitivity of the model has been tested in different ways.

2 Modelling

The water-steam system of the thermal power plant consists of several process components. System models can be built by combining model blocks of these sub-systems. In this work, the blocks can be classified as follows: pressure and flow values are calculated in the pressure point model blocks, and pressure rise and drop and heat transfer are modelled in the pump, valve and process element/heat exchanger model blocks.

2.1 Basic Process Components

Figure 1 presents a simple example of a process model, which contains two pressure points, a pump, valve and a process element. The water pressure and the outgoing flow are calculated in the point. Pump and valve models calculate pressure rise Δp_{pump} [Pa] and drop $\Delta p_{\mathrm{valve}}$ [Pa], respectively. Pressure rise is a function of the pump speed. Pressure drop between pressure points can be controlled by manipulating the valve stem position. In the literature there are several different equations to the pressure and flow calculation of pumps and valves (Kirmanen et al., 2012; Ordys et al., 1994). The input variables of all model blocks are mass flow, inlet pressure and inlet enthalpy of water. Similarly, the outputs are mass flow, outlet pressure and outlet enthalpy. Parameters of blocks can be determined based on the dimensions of the component and other properties.

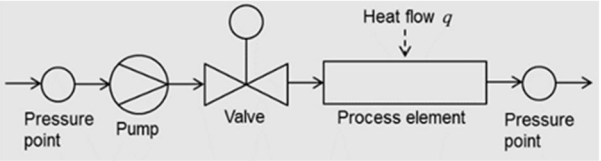


Figure 1. An example of a structure of the pressure and flow model.

Process element model block represents different water-steam side components such evaporators, superheaters, economisers and pipes. It is also possible to divide the modelled sub-processes into several model blocks because the model blocks can be connected together. The element model block contains calculations of pressure drop Δp_{elem} [Pa] according to (Pioro et al., 2004)

$$\Delta p_{\mathrm{elem}} = \Delta p_{\mathrm{fric}} + \Delta p_{\mathrm{loc}} + \Delta p_{\mathrm{acc}} + \Delta p_{\mathrm{elev}} \qquad (1)$$

where Δp_{fric}, Δp_{loc}, Δp_{acc} and Δp_{elev} are the pressure drops [Pa] due to frictional resistance, local flow obstruction, acceleration of flow and gravity. There are several equations in literature, which can be applied to pressure drop calculation. (Pioro et al., 2004; Tong and Tang, 1997)

The process element model block calculates also the time derivate of water enthalpy, which is affected by the enthalpy of inlet water flow and the heat transfer between water and walls of the process element. Moreover, the heat transfer and pressure drop depend on water properties and flow rate. The properties of water are interpolated from water-steam tables. The block is able to handle the thermodynamics of liquid water, saturated water-steam mixture, superheated steam, and supercritical fluid. The model block selects suitable water side heat transfer coefficient and pressure drop for different situations. The model block includes also metal walls (and possible refractory layers) of represented process components. Also the conductive heat transfer through these walls and layers has been modelled. The model block calculates time derivatives of metal and refractory walls temperatures. (Yli-Fossi et al., 2011; Yli-Fossi et al., 2012)

Pressure point model block calculates time derivatives of pressure and outlet fluid mass flow(s). The calculation of pressure p [Pa] for single-phase water and steam is based on the following equation (Lu, 1999)

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{(w_{\mathrm{i}}h_{\mathrm{i}} - w_{\mathrm{o}}h_{\mathrm{o}} + q) - \left(\frac{\rho}{\frac{\partial\rho}{\partial h}} + h\right)(w_{\mathrm{i}} - w_{\mathrm{o}})}{V\left(1 + \frac{\rho\frac{\partial\rho}{\partial p}}{\frac{\partial\rho}{\partial h}}\right)} \qquad (2)$$

where w_{i} and w_{o} are inlet and outlet mass flow rates [kg/s]. h_{i} and h_{o} are inlet and outlet enthalpies [J/kg]. q is heat flow [J/s] to fluid. ρ and h are density [kg/m^3] and enthalpy in a control volume V [m^3]. A certain part of the modelled process can be defined as the control volume.

The time derivative of pressure for two-phase liquid and vapour mixture can be calculated as (Lu, 1999)

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{(w_\mathrm{i}h_\mathrm{i} - w_\mathrm{o}h_\mathrm{o} + q) - \frac{\rho_\mathrm{l}h_\mathrm{l} - \rho_\mathrm{v}h_\mathrm{v}}{\rho_\mathrm{l} - \rho_\mathrm{v}}(w_\mathrm{i} - w_\mathrm{o})}{e_\mathrm{l}V_\mathrm{l} + \mathrm{e}_\mathrm{v}V_\mathrm{v}}$$

$$e_\mathrm{l} = \rho_\mathrm{l}\frac{\partial h_\mathrm{l}}{\partial p} + \frac{\rho_\mathrm{v}(h_\mathrm{v} - h_\mathrm{l})}{\rho_\mathrm{l} - \rho_\mathrm{v}}\frac{\partial \rho_\mathrm{l}}{\partial p} - 1$$

$$e_\mathrm{v} = \rho_\mathrm{v}\frac{\partial v}{\partial p} + \frac{\rho_\mathrm{l}(h_\mathrm{v} - h_\mathrm{l})}{\rho_\mathrm{l} - \rho_\mathrm{v}}\frac{\partial \rho_\mathrm{v}}{\partial p} - 1 \qquad (3)$$

where ρ_l and ρ_v are liquid and vapour densities [kg/m^3] and h_l and h_v are liquid and vapour enthalpies [J/kg] in the control volume V [m^3].

The time derivative of liquid volume V_l [m^3] in the control volume can be expressed as (Lu, 1999)

$$\frac{\mathrm{d}V_\mathrm{l}}{\mathrm{d}t} = \frac{w_\mathrm{i} - w_\mathrm{o} - \left(V_\mathrm{l}\frac{\partial \rho_\mathrm{l}}{\partial p} + V_\mathrm{v}\frac{\partial \rho_\mathrm{v}}{\partial p}\right)\frac{\mathrm{d}p}{\mathrm{d}t}}{\rho_\mathrm{l} - \rho_\mathrm{v}} \qquad (4)$$

Vapour volume V_l [m^3] in the control volume is (Lu, 1999)

$$V_\mathrm{v} = V - V_\mathrm{l} \qquad (5)$$

The time derivative of the outlet fluid mass flow [kg/s] from the pressure point model block is calculated as (Fabian, 2009)

$$\frac{\mathrm{d}w}{\mathrm{d}t} = \frac{p - p_\mathrm{next} - \Delta p_\mathrm{tot}}{L/A} \qquad (6)$$

where p is pressure [Pa] in the pressure point and p_next is pressure [Pa] in the next pressure point in the flow direction. According to Figure 1, p is the pressure of the left pressure point model block and p_next is the pressure of the right point. L and A are the total length [m] and inner cross-sectional areas [m^2] of tubes of process components between pressure point model blocks for p and p_next. Fluid inertia L/A is most significant in long and slender tubes. Δp_tot [Pa] is the total pressure drop between the pressure points. In case of Figure 1, Δp_tot can be determined as

$$p_\mathrm{tot} = \Delta p_\mathrm{pump} + \Delta p_\mathrm{valve} + \Delta p_\mathrm{elem} \qquad (7)$$

In literature there are alternative ways to calculate mass flows (Jäntti, 1996; Lu, 1999). In this work the Equation 6, based on Newton's second law, is applied. The same equation is used for liquid water and vapour. The compressibility of vapour is considered in the pressure equations.

2.2 Boiler Evaporator Loop

Steam boilers can be classified according to the structures of the evaporation process. The types are drum boilers with natural or forced circulation, and a once-through design. Boiling occurs in evaporator tubes which form furnace walls. Figure 2 illustrates the structure of the forced circulation system. The main components of the system are: the steam drum, downcomer pipes, circulation pump and riser tubes. The drum separates steam from the saturated water-steam mixture flowing out from the riser tubes. Downcomer pipes convey water from the drum to the lowest part of the evaporator under the furnace floor. Riser tubes lead water back from the bottom of the furnace to the drum. The outlet flow from the riser tubes consist of saturated water and saturated steam, which are separated from each other in the steam drum. Saturated water mixed with new feed water is circulated back from the drum to the evaporator and saturated steam is led to the superheaters to be heated up. In a natural circulation boiler, the density difference between liquid water in downcomer and water-steam mixture in riser provides the pressure difference required for circulation. In a forced circulation boiler, a pump is used to increase the circulation rate as compared to that of a natural circulation boiler. (Åström and Bell, 2000; Li et al., 2005)

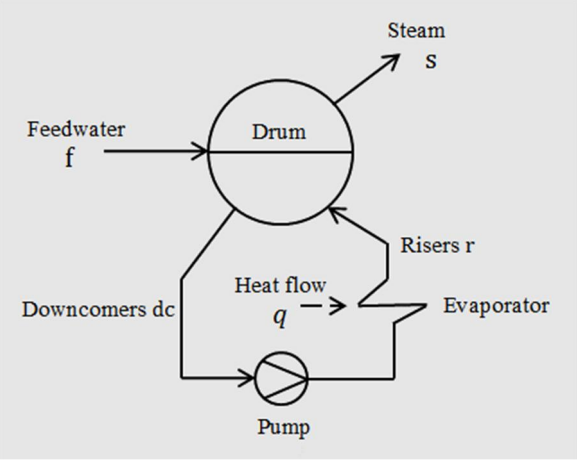


Figure 2. The structure of the circulating evaporation system.

Several studies about the modelling of the circulating evaporation process can be found in the literature. In the presented model the drum pressure is determined as (Åström and Bell, 2000)

$$\frac{\mathrm{d}p}{\mathrm{d}t} =$$

$$\frac{(w_\mathrm{f}h_\mathrm{f} - w_\mathrm{s}h_\mathrm{s} + q) - (\rho_\mathrm{l}h_\mathrm{l} - \rho_\mathrm{v}h_\mathrm{v})\frac{\mathrm{d}V_\mathrm{lt}}{\mathrm{d}t}}{V_\mathrm{lt}\left(h_\mathrm{l}\frac{\partial\rho_\mathrm{l}}{\partial p} + \rho_\mathrm{l}\frac{\partial h_\mathrm{l}}{\partial p}\right) + V_\mathrm{vt}\left(h_\mathrm{v}\frac{\partial\rho_\mathrm{v}}{\partial p} + \rho_\mathrm{v}\frac{\partial h_\mathrm{v}}{\partial p}\right) - V_\mathrm{t} + m_\mathrm{t}C_p\frac{\partial T_\mathrm{s}}{\partial p}}$$

$$\frac{\mathrm{d}V_\mathrm{lt}}{\mathrm{d}t} = \frac{w_\mathrm{f} - w_\mathrm{s} - \left(V_\mathrm{lt}\frac{\partial\rho_\mathrm{l}}{\partial p} + V_\mathrm{vt}\frac{\partial\rho_\mathrm{v}}{\partial p}\right)\frac{\mathrm{d}p}{\mathrm{d}t}}{\rho_\mathrm{l} - \rho_\mathrm{v}} \tag{8}$$

where w_f and w_s are the mass flows [kg/s] of the feed water to the drum and the steam flow from the drum. h_f and h_s are the enthalpies [J/kg] of the feedwater and the saturated steam. ρ_l and ρ_v are liquid and vapour densities [kg/m^3] and h_l and h_v are liquid and vapour enthalpies [J/kg] in the drum. V_lt and V_vt are liquid and steam volumes [m^3] in the total circulation evaporation system. m_t is the mass [kg] of the metal structure of the system. C_p is the specific heat capacity [J/(K*kg)] of metal and T_s is the saturation temperature [K] of the steam.

The flow through the downcomers can be calculated as (Li et al., 2005)

$$\frac{\mathrm{d}w_\mathrm{dc}}{\mathrm{d}t} = \frac{\rho_\mathrm{dc}gH_\mathrm{dc} - \int_0^{L_\mathrm{r}}\rho_\mathrm{r}\,g\,\mathrm{d}z - \Delta p_\mathrm{r} - \Delta p_\mathrm{dc} + \Delta p_\mathrm{pump}}{(L_\mathrm{dc} + L_\mathrm{r})/A_\mathrm{dc}} \tag{9}$$

where ρ_dc and ρ_r are the densities [kg/m^3] of the process fluid in the downcomers and the risers. L_dc and L_r are the the lengths [m] of of the downcomers and the risers. H_dc is the height [m] of the downcomers. z is the vertical position [m] in the risers. g is the gravitation constant [m/s^2]. A_dc is the inner cross-sectional area [m^2] of the risers. Δp_dc, Δp_r and Δp_pump are the pressure drops [Pa] across the downcomers, the risers and the circulation pump. Δp_pump is omitted if a natural circulation boiler is modelled.

The flow through the risers can be presented as (Åström and Bell, 2000)

$$\begin{aligned} w_\mathrm{r} = w_\mathrm{dc} - V_\mathrm{r}\Big(&\alpha_v\frac{\partial\rho_\mathrm{v}}{\partial p} + \left(1 - \alpha_v\frac{\partial\rho_\mathrm{l}}{\partial p}\right) \\ &+ (\rho_\mathrm{l} - \rho_\mathrm{v})\frac{\partial\alpha_v}{\partial p}\Big)\frac{\mathrm{d}p}{\mathrm{d}t} \\ &+ (\rho_\mathrm{l} - \rho_\mathrm{v})V_\mathrm{r}\frac{\partial\alpha_v}{\partial\alpha_r}\frac{\mathrm{d}\alpha_r}{\mathrm{d}t} \end{aligned} \tag{10}$$

where V_r is the volume [m^3] of the risers. α_v is the average steam volume fraction. α_r is steam mass fraction at the risers' outlet.

Condensate and steam flow rates through the surface level in the drum can also be modelled by the equations found in the literature (Åström and Bell, 2000).

2.3 Test Model

This work is a part of the wider modelling work where different types of steam power plants have been modelled. These models include also the air-flue gas system and the dynamics of combustion and the heat transfer from hot flue gases to heat exchanger structures. Also the main control loops are included. The models can be used for several purposes, such as control design and process development. The focus of this paper is on the testing of pressure and flow calculations. A fairly simple model is used for the testing, because it facilitates the analysis of the computation and the examination of the functionality. The dynamic water-steam system model has been developed using Simulink and Matlab by The MathWorks Inc. The process is modelled as a continuous time model.

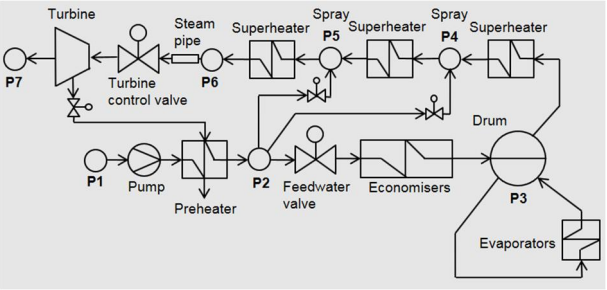


Figure 3. The diagram of the dynamic power plant water-steam system model.

Figure 3 illustrates the structure of the test model. The water-steam system model consists of a feedwater pump and a valve, a preheater, economisers, a drum, and a natural circulation evaporator, three superheaters, two attemperation sprays and a steam pipe, a turbine valve and a steam turbine. The flue gas side has not been included in this model. Suitable heat flows are only added to the model blocks of the economisers, the evaporators and the superheaters as input variables. Feed water is heated by the preheater taking heat energy from steam extracted from the steam turbine. The economisers (flue gas preheaters) increase feed water temperature after the preheaters near to the boiling point. The evaporators generate steam and the drum separates steam from saturated liquid-vapour mixture. The superheaters increase the live steam temperature before the turbine. Superheated steam temperature is controlled by attemperation sprays. The cooling water is taken before the feedwater valve. The spray valves are

also modelled. The main pressure is controlled by the heat power directed to the evaporators. The turbine control valve is used to regulate the steam flow to the turbine. A drum level control is also implemented in the model. The level is adjusted with the feedwater flow. The controllers are needed to achieve the desired steady state of the system. The pressure point model blocks have been marked on Figure 3 with P1…P7. The presented equations are applied in each blocks.

3 Simulation Tests

The presented model has been tested in different ways. The numerical solvers were compared with test runs. The effect of the correctness of the initial values of the states has been studied. The tolerance of the model to the changes in the parameters and the structure was also examined.

3.1 Comparison of Numerical Solvers

The MATLAB offers several solvers for solving ordinary differential equations (ODEs). The ode45 solver was chosen as a reference solver which is more suitable for nonstiff systems. The ode45 solver uses a variable step and one-step Runge-Kutta procedure. It calculates both the fourth and fifth order approximations. The MATLAB/Simulink environment contains four solvers, ode15s, ode23s, ode23t and ode23tb, all of which are designed to solve stiff equations. The numbers in the names represent the orders of the approximations. Ode15s solver turned out to be very slow for the test model and for this reason it was left out from the more exact study. The ode15 is a multistep solver which is based on the numerical differentiation formulas. The one-step ode23s solver is based on a modified Rosenbrock formula of the second order. The ode23t is an implementation of the trapezoidal rule using a free interpolation. The ode23tb is an implementation of TR-BDF2. (MATLAB: Documentation, 2016)

The comparative simulation test runs were performed by different solvers. Each solver was tested with three values, 1e-3, 1e-5 and 1e-8, of the relative tolerance. The relative tolerance measures the error relative to the size of each state. The relative tolerance represents a percentage of the state's value. For example, value 1e-3, means that the computed state will be accurate to within 0.1%. The parameter of the absolute tolerance was set in auto-mode. The same initial values of the states were set for each simulation run. The model was near steady state at full load. The lengths of the runs were 100 simulated seconds. The results of the simulation runs are summarized in Table 1. The speed is presented as a ratio between a real time and consumed simulation time. Decrement of the tolerance decreases simulation speed and noise of calculated variables. The solvers ode45 and ode23s are clearly slower than the solvers ode23t and ode23tb. Figures 4 and 5 present simulated steam pressure and mass flow in the point P6 after the last superheater. Oscillations of the simulated results are distinctly seen especially when the relative tolerance is 1e-3. The solver is ode23tb, which seems the most efficient on the basis of Table 1. In this case, it seems that the suitable compromise between simulation speed and accuracy is achieved with relative tolerance 1e-5.

Table 1. Comparison of the Solvers

Name	*Relative tolerance*	*P6: Pressure deviation: variance [%]*	*P6: Mass flow deviation: variance [%]*	*Simulation speed: real time / simulated time*
ode45	1e-3	0.1075e-2	1.1419e-2	0.4463
	1e-5	0.0184e-2	0.3303e-2	0.4244
	1e-8	0.0185e-2	0.3294e-2	0.3497
ode15s	1e-3	not tested	not tested	very slow
ode23s	1e-3	0.0413e-2	0.4899e-2	0.5520
	1e-5	0.0414e-2	0.4905e-2	0.5428
	1e-8	0.0324e-2	0.4291e-2	0.3042
ode23t	1e-3	0.0261e-2	0.3646e-2	7.3868
	1e-5	0.0223e-2	0.3571e-2	4.8860
	1e-8	0.0220e-2	0.3545e-2	1.0878
ode23tb	1e-3	0.7236e-2	3.8129e-2	8.3178
	1e-5	0.0516e-2	0.4356e-2	8.1136
	1e-8	0.0222e-2	0.3566e-2	3.3905

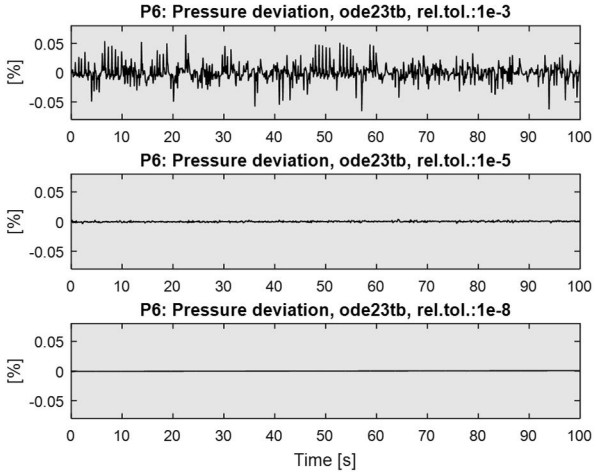


Figure 4. Simulated steam pressure in the point P6 using different values of the relative tolerance by the solver ode23tb.

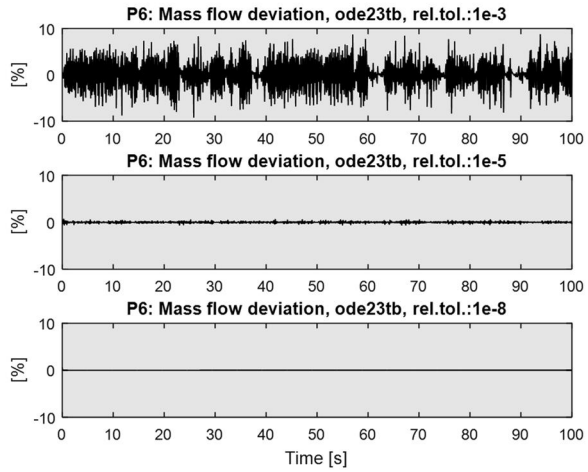


Figure 5. Simulated steam mass flow in the point P6 using different values of the relative tolerance by the solver ode23tb**.**

3.2 Effects of Initial State and Parameters

The initial values of the model state have high significance from the point of view of the functionality of the model. To setting the model first time to the steady state may be difficult. The correct initial values can be calculated or estimated before the first simulation run. However, this can be time-consuming and troublesome. The second alternative is to simulate the model to the desired state. The challenge is that wrong initial values can lead the simulation run to fail. In any case, when the desired state has been reached, it can be saved. Saved states can be used to the initialisation of the state values. The changing of the parameters and structure of the model may also cause problems. For this reason, it is a significant advantage if the model tolerates different state values and updating of the model.

Figure 6 represents a case where the initial values of the outlet pressure and outlet mass flow of the pressure point P5 were set incorrect in the full load state. The deviation of the outlet pressure was set -15 % and the outlet mass flow -100 % (0 kg/s) before a simulation run. Other model state variables were at the proper values. All controllers were set on manual mode. The figure proves that the model tolerates the false initial values and settles in steady state. However, if the deviation of the pressure was set -20 %, then the simulation run failed.

The drum model functionality also was tested in the same way. The initial value of the drum pressure was set about -15 % smaller than in the full load situation. Figure 7 shows that the drum pressure began to rise immediately at the beginning of the simulation. The steam flow from the drum was decreased because the drum pressure was smaller than in the next pressure point P4. Steam flow returned normal when the drum pressure rose enough. Figure 8 presents the mass flows of the downcomers and the risers in the evaporation loop in the same simulation run. The pressure deviation affected the riser flow more strongly than into the dowcomer flow. At the end of the simulation the flows reached the balances. The model tolerated the change of the initial value of the drum pressure.

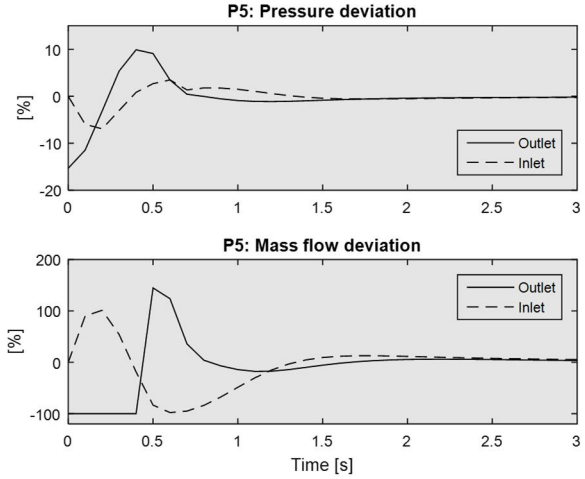


Figure 6. Simulated pressures and mass flows in the point P5.

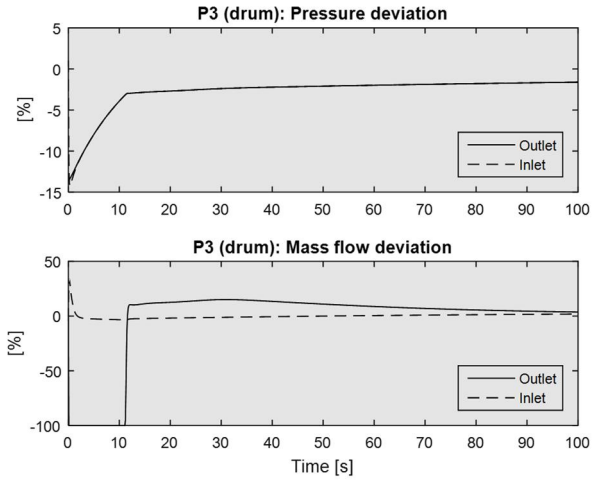


Figure 7. Simulated pressures and mass flows in the drum.

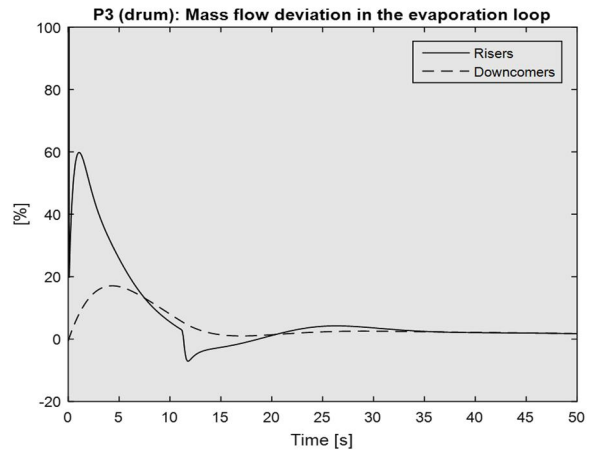


Figure 8. Simulated downcomer and riser mass flow in the drum.

Effect of the changes in the model parameters was also tested by the model. It turned out that the model tolerated well a change in the size of the drum and other process components. Effect of the editing of the structure of the model also was studied. A new superheater model was added between the present superheater and the pressure point P4. Thus, the new model version contained two superheater model blocks between the drum and P4. The controllers were set to auto-mode in this simulation test. The adding of the model block caused oscillation before the model settled in the balance, which can be seen in Figure 9. The initial state of the new block was far from the reasonable values. For example, in the start situation the vapour mass fraction was zero inside the new superheater. In spite of this the model was able to handle the unconventional situation. The reasonable values were reached in the end of the simulation run.

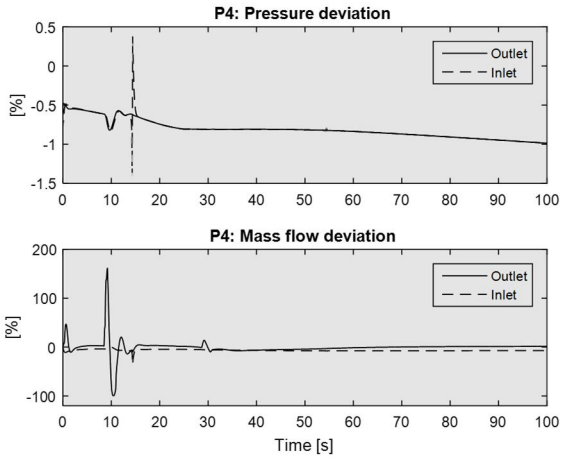


Figure 9. Simulated pressures and mass flows in the point P4.

4 Conclusions

A dynamic power plant water side model and selected water pressure and flow equations has been presented. The equations are based on physical equations. The suitable solver's choice was also studied and tested. Functionality testing proved that the model tolerates different changes well within certain limits. A workable way to make flexible models with pressure and flow calculation was found.

Acknowledgements

This work was carried out in the FLEXe research program coordinated by CLIC Innovation Ltd. The work has also funded Neles 30-year Anniversary Foundation. These supports are gratefully acknowledged.

References

B. Fabian, *Analytical System Dynamics: Modeling and Simulation*. Springer Science+Business Media, 2009.

T. Jäntti, *Dynamic simulation of a circulating fluidized bed boiler*. Master of Science Thesis, Lappeenranta University of Technology, 1996.

J. Kirmanen, I. Niemelä, J. Pyötsiä, M. Simula, M. Hauhia, J. Riihilahti, V. Lempinen, Koukkuluoma J. and P. Kanerva, *Flow Control Manual*. Metso Automation Inc., 6th Edition, 2011.

B. Li, T. Chen and D. Yang, DBSSP – A computer program for simulation of controlled circulation boiler and natural circulation boiler start up behavior. *Energy Conversion and Management*, 46:533-549, 2005.

S. Lu, Dynamic modelling and simulation of power plant systems. *In Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*, 213(1):7-22, 1999.

MATLAB: Documentation. The MathWorks Inc., 2016.

E. Majuri, *Flow Distribution and Stability in a Two-phase Natural Circulation System*. Master of Science Thesis, Tampere University of Technology, 2012.

A. Ordys, A.W. Pike, M.A. Johnson, R.M. Katebi and M.J. Grimble, *Modelling and Simulation of Power Generation Plants*. Springer-Verlag London Limited, 1st Edition, 1994

I.L. Pioro, R.B. Duffey and T.J. Dumouchel, Hydraulic resistance of fluids flowing in channels at supercritical pressures (survey). *Nuclear Engineer and Design*, 231:187-197, 2004.

L.S. Tong and Y.S. Tang, *Boiling Heat Transfer and Two-Phase Flow*. 2nd edition, Taylor & Francis, 1997.

T. Yli-Fossi, P. Köykkä and Y. Majanne, A Generalized Dynamic Water Side Model for a Once-Through Benson Boiler. *In Preprints of the 18th IFAC World Congress*: 7049-7054, Milano, Italy, 28 August-2 September 2011.

T. Yli-Fossi, P. Köykkä and Y. Majanne, A Tuning Tool for Gas Side Heat Transfer Coefficients of a Boiler Model. *In Proceedings of the 8th IFAC Power Plant and Power System Control Symposium*: 180-185, Toulouse, France, 2-5 September 2012.

R. Åman, *Methods and Models for Accelerating Dynamic Simulation of Fluid Power Circu*its. Dissertation, Lappeenranta University of Technology, 2011.

K.J. Åström and R.D. Bell, Drum-boiler dynamics. *Automatica* 36:363-378, 2000.

Mathematical Modeling of the Parabolic Trough Collector Field of the TCP-100 Research Plant

Antonio J. Gallego[1] Luis J. Yebra[2,3] Eduardo F. Camacho[1] Adolfo J. Sánchez[1]

[1]Dpto. de Ingeniería de Sistemas y Automática, Universidad de Sevilla, Spain,
{antgallen@gmail.com,eduardo@esi.us.es, adolfo.j.sanchez@ieee.org}
[2]Plataforma Solar de Almería, CIEMAT, Spain, luis.yebra@psa.es
[3]CIESOL, Joint Centre of the University of Almería-CIEMAT, Spain

Abstract

There are two main drawbacks when operating solar energy systems: a) the resulting energy costs are not yet competitive and b) solar energy is not always available when needed. In order to improve the overall solar plants efficiency, advances control techniques play an important role. In order to develop efficient and robust control techniques, the use of accurate mathematical models is crucial. In this paper, the mathematical modeling of the new TCP-100 parabolic trough collector (PTC) research facility at the Plataforma Solar de Almería is presented. Some simulations are shown to demonstrate the adequate behavior of the model compared to the facility design conditions.

Keywords: solar energy, parabolic trough collector, modeling, simulation

1 Introduction

The interest in renewable energy sources such as solar energy, experienced a great impulse after the Big Oil Crisis in the 70s. Driven mainly by economic factors, this interest decreased when oil prices fell. Nowadays, there is a renewed interest in renewable energies spurred by the need of reducing the environmental impact produced by the use of fossil energy systems ((Goswami et al., 2000) (Camacho and Berenguel, 2012)). Solar energy is, by far, the most abundant source of renewable energy (IRENA, 2012).

Many solar electricity production, furnaces, heating and solar cooling systems have been developed in the last decade (Camacho et al., 2012). The main technologies for converting solar energy into electricity are photovoltaic (PV) and concentrated solar thermal (CST). Parabolic trough, solar towers, Fresnel collector and solar dishes are the most used technologies for concentrating solar energy.

As example of the above mentioned, we can mention the following commercial solar plants: The 9 SEGS trough plants (354 MW) which commissioned between 1985 and 1990 in California, are considered to be the first commercial plants. Most of the commercial solar plants have been built and commissioned in the last decade. As examples we can mention the three 50 MW parabolic trough plants Andasol 1, 2 and 3 in Guadix (Spain), the solar tower plants of Abengoa PS10 and PS20, Gemasolar solar tower built by Torresol Energy, the three 50 MW Solnova and the two 50 MW Helioenery parabolic trough plants of Abengoa in Spain, and the SOLANA and Mojave Solar parabolic trough plant now operating in Arizona, of 280 MW power production each (Camacho and Gallego, 2013).

One of the first experimental solar trough plants was the solar field ACUREX at the Plataforma Solar de Almería (PSA). It consisted of a field of solar collectors, a heat storage system and an electrical conversion unit (0.5 MW Stal-Laval turbine). This plant has been operating from 1980 to 2013, and many control strategies have been tested there ((Rubio et al., 2006) (Lemos et al., 2000) (Berenguel, 1996) (Gallego et al., 2013)).

There are two main drawbacks when operating solar energy systems: a) the resulting energy costs are not yet competitive and b) solar energy is not always available when needed. Considerable research efforts are being devoted to develop techniques which may help to overcome these drawbacks (Camacho et al., 2011); advanced control is one of those techniques which can help on reduce operating costs and increase solar plants performance (Camacho and Gallego, 2015).

In order to develop control and optimization algorithms for solar energy systems, obtaining an accurate dynamic model is very useful. This paper presents a mathematical model of the new PTC TCP-100 research facility at the PSA, currently under construction.

A considerable research effort has been done in the past concerning the developing of accurate mathematical models describing the dynamics of parabolic trough systems. One of the first work describing the equations which govern the behavior of a parabolic trough loop was done in (Carmona, 1985). For example, in (Yebra et al., 2006), an object oriented modeling and simulation of parabolic trough collectors with modelica is presented. In (Yilmaz and Soylemez, 2014), a complex analysis based on solar, optical and thermal models is developed by using differential and non-linear algebraic correlations. In (JAI and CHALQI, 2013), a mathematical model that describes the heat exchange between the main components of a thermal solar collector in an Integrated Solar Combined Cy-

cle (ISCC) plant is described. More precise and complex models dealing with modeling of parabolic trough plants with direct vapor generation are described in (Bonilla et al., 2012), (Bonilla, 2013), (yeb, 2005) and (Bonilla et al., 2011).

A PTC solar plant consists mainly of: a collector field, a power conversion system (PCS), a storage system and auxiliary elements such as pumps, pipes and valves ((Duffie and Beckman, 1991), (Camacho et al., 2013)). The solar collector field is formed by PTCs that collect solar radiation and focus it onto a tube in which a heat transfer fluid, usually synthetic oils, circulates. As the oil passes through the metal tube, it is heated up and then used by the PCS to produce electricity by means of a turbine. The storage system is necessary to cover possible mismatches between the solar energy available and the demand. This is one of the advantages of solar thermal energy: the storage of the thermal energy is easier and cheaper than the storage of electrical energy (Herrmann and Kearney, 2002).

The new TCP-100 has new features compared to the ACUREX solar field: the solar tracking system is North-South axis instead of West-East axis of the ACUREX field (Camacho et al., 2007). The collectors and metal tubes are greater and the working temperatures are about 350-380 °C whereas the normal temperatures of the ACUREX field were about 250-280 °C (Camacho et al., 1997). A more complete description of the TCP-100 research facility solar field is carried out in section 2.

The model uses data from data sheets of components and designing conditions from technical documentation of this facility. In particular, parameters related to collectors' size, diameter of the metal tube and the overall optical efficiency. The main characteristics of the heat transfer fluid have been obtained using data from the datasheet provided by the supplier (Dowtherm). Thermal losses will be obtained more precisely when experimental data is available. For simulation purposes, data from the provider of the PTCs is used as a first approximation.

The paper is organized as follows: section 2 provides a complete description of the TCP-100 PTC solar field. Section 3 describes the main assumptions and equations of the mathematical model. Section 4 shows simulations describing the behavior of the model. Finally, section 5 draws to a close with some conclusions.

2 TCP-100 solar field description

A new parabolic trough collector facility has been erected at Plataforma Solar de Almería (CIEMAT), in replacement of the so many times referenced ACUREX field that had been operated for more than 30 years. The new facility is named TCP-100 and has been designed mainly to develop automatic control algorithms for parabolic trough solar fields .

The TCP-100 solar field is formed by three loops of parabolic trough collectors (PTC), each of them composed by two PTCs in a North-South orientation. Each of 6 PTCs is 100 m length, formed by 8 modules in parallel. Fig. 1 shows the first PTC belonging to the first loop.

Figure 1. Lateral view of the first TCP-100 PTC in the first loop at Plataforma Solar de Almería (PSA-CIEMAT). It is composed of 8 modules of 12 meters length each.

Figure 2. Top view of the TCP-100 field at Plataforma Solar de Almería (PSA-CIEMAT). The three loops are shown, with two PTCs in each of them, numbered from 1 (rightmost) to 6 (leftmost). The first loop is formed by the connected pair 1-2 (right loop), the second loop by 3-4 (center loop) and the third by 5-6 (left loop).

The PTCs in each loop are connected in the South extreme, and the *colder* PTC will be always the first in the row, placed at the right part (see Fig. 2).

Remarkable features of the new solar field are those aimed at the experimentation of advanced control techniques, with an important quantity of sensors and actuators with respect to its predecessor ACUREX. These features are summed up:

- Inlet and outlet solar field temperature sensors.
- For each loop, inlet and outlet temperatures are measured. Inside the loop, for each PTC: inlet, outlet and middle point temperatures sensors are available.
- Volumetric flow rate for each loop.
- The aperture of control valves at the input of each loop can be controlled.

The heat transfer fluid is Syltherm 800, suitable for the operating conditions of this new field. Although not treated in this paper, it is worth mentioning that this new solar field is connected to a thermocline storage tank and a cooler cycle through a heat exchanger. The operating

conditions for a nominal solar radiation of 900 $\mathrm{W/m^2}$ are: $\dot{m} = 18.72$ kg/s, inlet and outlet temperatures of 330 °C and 380 °C respectively. A schematic diagram of the field is shown in Fig. 3.

3 Mathematical modeling of TCP-100 solar field

Each of the TCP-100 loops consists of two eight module PTCs suitably connected in series. Each collector measures 100 m long and the passive parts joining them (parts where solar radiation does not reach the tube) measures 24 m long.

This sort of systems can be modeled by using a lumped description (concentrated parameter model) or by a distributed parameter model (Gallego and Camacho, 2012). The approach used here is the distributed parameter model for each loop. The whole solar, composed of the three parallel loops, can be modeled by adding loops in parallel.

The model equations are the same used in the ACUREX solar field developed in (Carmona, 1985) and (Camacho et al., 1997). The model consists of the following system of non-linear partial differential equations (PDE) describing the energy balance:

$$\rho_m C_m A_m \frac{\partial T_m}{\partial t} = IK_{opt} cos(\theta) G - H_l G(T_m - Ta) - LH_t(T_m - T_f) \quad (1)$$

$$\rho_f C_f A_f \frac{\partial T_f}{\partial t} + \rho_f C_f q \frac{\partial T_f}{\partial x} = LH_t(T_m - T_f) \quad (2)$$

Where the subindex m refers to metal and f refers to the fluid. The model parameters and their units are shown in Table 1.

In the following, all the parameters needed for the model are described, with the exception of geometric parameters (lengths and areas). The metal density and specific heat correspond to stainless steel tube 321.

3.1 Optical and geometric efficiencies

The optical efficiency, K_{opt}, takes into account elements such as reflectivity, absorptance, interception factor and others. According to the maker, the peak optical efficiency is about 0.76.

The geometric efficiency, $cos(\theta)$, is determined by the position of the mirrors with respect to the radiation beam vector. It depends on hourly angle, solar hour, declination, Julianne day, local latitude and collector dimensions (Goswami et al., 2000).

Parabolic trough collectors usually track the sun with one degree of freedom using the E-W axis (as the ACUREX field did) or the N-S axis (Oden and Abu-Mulaweh, 2013). Solar tracking maintains the plane of a solar beam so that it is always normal to the collector aperture. Commercial plants use the N-S axis tracking, because it improves greatly the amount of direct solar radiation collected compared to the E-W axis tracking throughout the year (Oden and Abu-Mulaweh, 2013). As stated previously, the TCP-100 solar field uses N-S axis tracking, and the expression for computing the $cos(\theta)$, is as follows ((Duffie and Beckman, 1991) (Osterholm and Palsson, 2014)).

Table 1. Parameters description

Symbol	*Description*	*Units*
t	Time	s
x	Space	m
ρ	Density	Kgm^{-3}
C	Specific heat capacity	$JK^{-1}kg^{-1}$
A	Cross Sectional Area	m^2
$T(x,y)$	Temperature	K,°C
$q(t)$	Oil flow rate	m^3s^{-1}
$I(t)$	Solar Radiation	Wm^{-2}
$cos(\theta)$	geometric efficiency	Unitless
K_{opt}	Optical efficiency	Unitless
G	Collector Aperture	m
$T_a(t)$	Ambient Temperature	K,°C
H_l	Global coefficient of thermal loss	Wm^{-2}°C^{-1}
H_t	Coefficient of heat transmission metal-fluid	Wm^{-2}°C^{-1}
L	Length of pipe line	m

$$cos(\theta) = \sqrt{cos(\theta_z)^2 + cos(\delta)^2 sin(\omega)^2} \quad (3)$$

Where δ is the declination and it can be obtained using the well known Spencer formulas by using equation (4), (Camacho et al., 2012).

$$\delta = 0.006918 - 0.399912 \cdot cos(\omega) + 0.070257 \cdot sin(\omega) - 0.006758 \cdot cos(2 \cdot \omega) + 0.000907 \cdot sin(2 \cdot \omega) - 0.002697 \cdot cos(3 \cdot \omega) + 0.00148 \cdot sin(3 \cdot \omega) \quad (4)$$

The variable ω, is the hourly angle, the angular displacement of the sun east or west of the local meridian due to rotation of the earth on its axis at 15 degrees per hour. It can be calculated as follows:

$$\omega = (T_{sun} - 12) \cdot 15 \cdot (\Pi/180) \quad (5)$$

Where T_{sun} is the solar hour (Blanco and Santigosa, 2017) which can be computed using the local time as follows (Osterholm and Palsson, 2014):

$$T_s = localhour + 4(L_{st} - L_{loc}) + E_t$$

Where L_{st} is the standard meridian for the local time zone, L_{loc} is the longitude of the location in degrees west

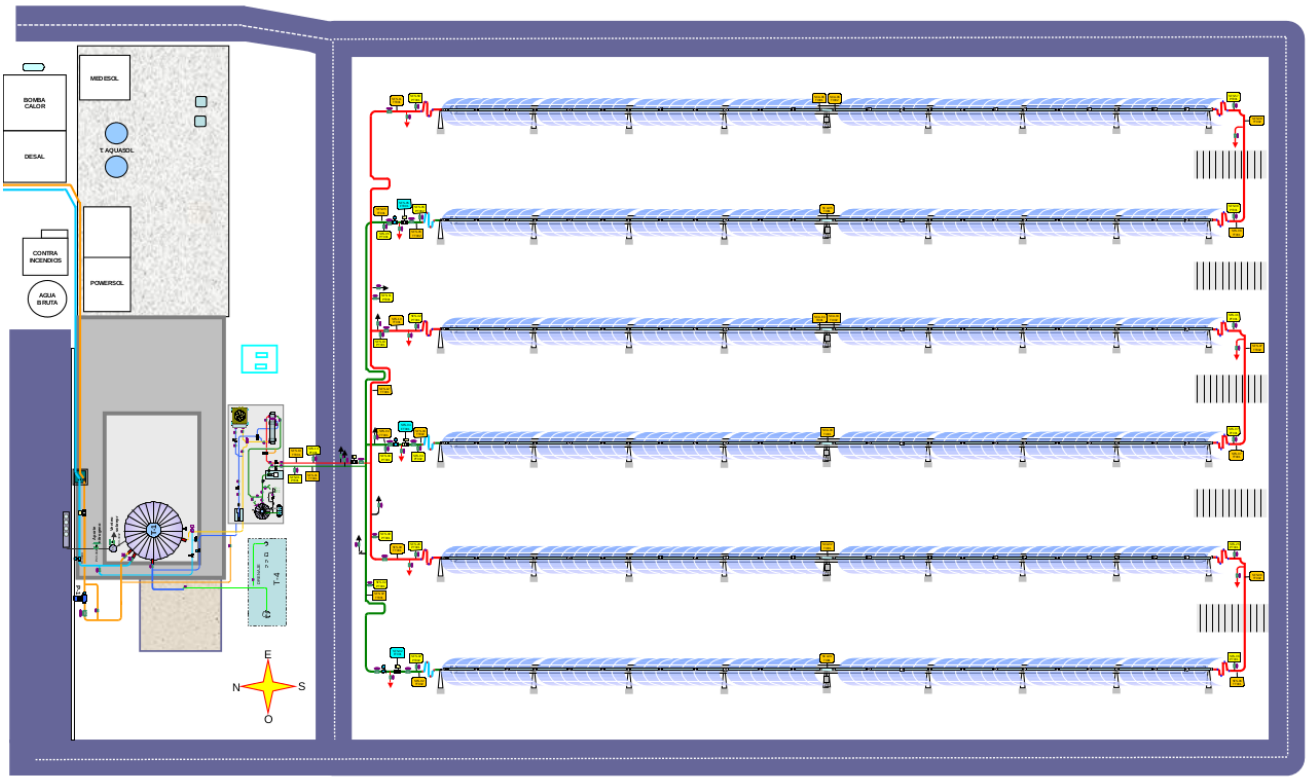


Figure 3. Schematic diagram of the TCP-100 solar field in which the three loops configuration.

and E_t is the equation of time (in minutes) which can be obtained using the following equation (6):

$$E_t = (0.000075 + 0.001868 \cdot cos(B) - 0.032077 \cdot sin(B) - 0.014615 \cdot cos(2 \cdot B) - 0.04089 \cdot sin(2 \cdot B)) \cdot 229.18 \quad (6)$$

With B computed as $B : (2 * \pi/365) * (JD - 1)$, where JD is the Julianne day.

The zenith angle is denoted by θ_z,the angle of incidence of beam radiation on a horizontal surface (Duffie and Beckman, 1991). It can be obtained by using the expression (7):

$$\theta_z = cos(\phi)cos(\delta)cos(\omega) + sin(\phi)sin(\delta) \quad (7)$$

The variable ϕ stands for the Latitude.

3.2 Characteristics of the heat transfer fluid

As stated before, the plant uses Syltherm 800 as a heat transfer fluid (HTF). The HTF is a highly stable, long-lasting silicone fluid designed for high temperature liquid phase operation. It can operate from -40 °C to 400 °C, without degradation.

The density ρ, specific heat C, and the coefficient of heat transmission have been obtained by data provided in the product data sheet. A polynomial adjustment was performed in order to obtain an expression valid for the entire range of operating temperatures. The expression obtained are:

$$\rho_f = -0.00048098T_f^2 - 0.811T_f + 953.65 \ (kg/m^3)$$

$$C_f = 0.0000001561T_f^2 + 1.70711T_f + 1574.2795 \ (J/(kg\ ^\circ C))$$

The coefficient of heat transmission has two parts: one depends on the temperature of the fluid and the other depends on the oil flow (Camacho et al., 1997). Obtaining the expression of this coefficient involves using complex convection heat transmission formulas (Baerh, 1965).

The expressions are as follows:

$$Hv(T) = 2 \cdot (-0.00016213T_f^3 + 1.221T_f^3 + 115.9983T_f + 12659.697$$
$$H_t = Hv(T)q^{0.8} \ (W/(m^2\ ^\circ C)) \quad (8)$$

3.3 Thermal losses

The thermal losses coefficient has to be obtained by using experimental data from the actual solar field. However, the field is not operative yet so the experimental data is not available. For simulation purposes, the coefficient of thermal losses has been considered to be similar to that

used in the ACUREX field, but taking into account that the overall thermal losses for 400 °C are about 265 W/m^2 as stated in the metal tube data sheet. The coefficient of thermal losses has the following expression:

$$H_l = 0.000357346 \cdot (T_m - T_a) - 0.00878632\ (W/(m^2\ K)) \quad (9)$$

When experimental data is available, this coefficient will be adjusted better.

4 Simulations

In this section, some simulation results are shown. In these simulations, typical operating values of solar radiation, inlet temperature and oil flow are used. These values are: inlet temperature: 330°C, outlet temperature: 380°C and the nominal mass flow is 18.72 kg/s for the whole solar field.

Figure 4 shows a simulation carried out on a summer day (Julianne day: 196). The upper part of figure 4 depicts the field temperatures (inlet and outlet temperatures). The bottom part of figure 4 shows the solar radiation (IDN), the modified solar radiation (mod IDN), namely, the product $I * cos(\theta)$, and the oil flow multiplied by 100 (kg/s). As can be seen, for the nominal operating conditions, the outlet temperature of the model is very close to that expected in the designing conditions (380 °C).

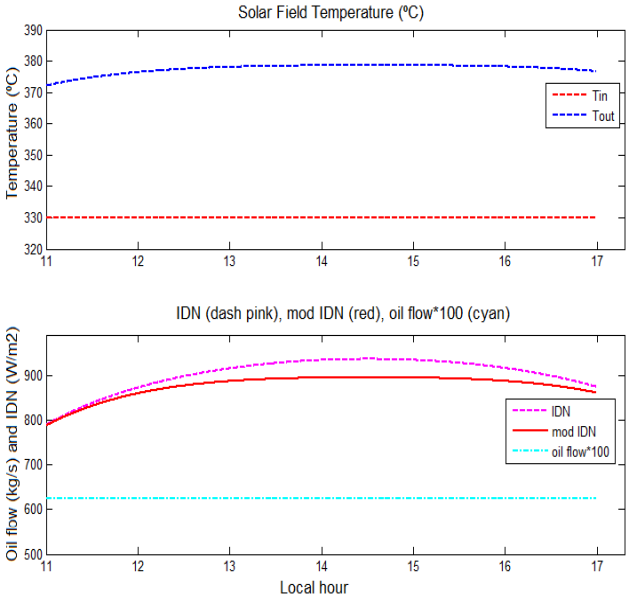


Figure 4. Simulation of a summer day.

Figure 5 plots a simulation performed with data from a winter day (Julianne day: 349). The outlet temperature is substantially lower because the modified solar radiation is much lower. This effect is produced by the $cos(\theta)$, which is smaller in winter months than summer months. In order to rise the outlet temperature, decreasing the oil flow is indispensable.

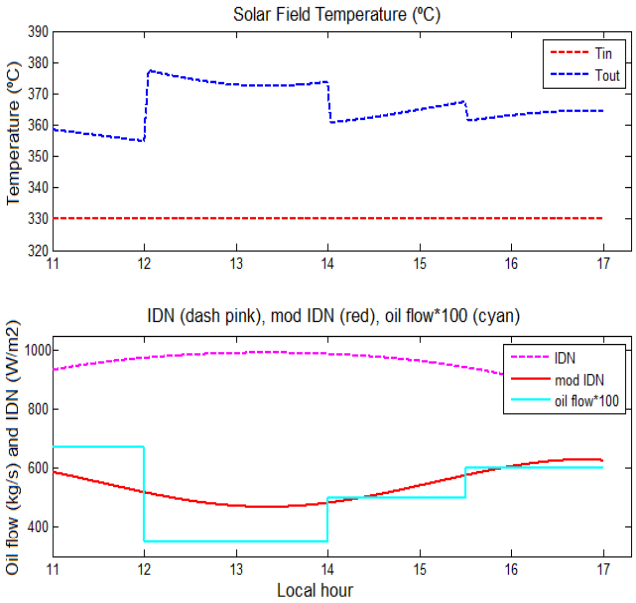


Figure 5. Simulation of a winter day.

The variation of the cosine of the incidence angle is shown in figure 6 for different months. It is shown that the variation of the $cos(\theta)$ is more pronounced throughout the day in winter days than in summer and spring days, with a deeper valley at the solar noon. In summer days, such as the one belonging to June, the cosine variation is smoother.

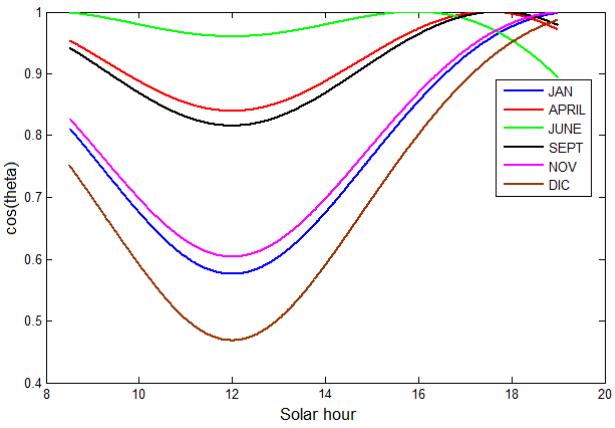


Figure 6. $cos(\theta)$ variation for different months.

Finally, a day with some transients in solar radiation, produced by scattered clouds, is simulated in figure 7. The clouds produce that the solar radiation decreased from 14.1 to 14.6 h approximately, and the solar field temperature decreases correspondingly. Some steps in the oil flow have been produced throughout the day, in order to maintain the outlet temperature around 365 °C, although when the passing clouds affect the solar field, the outlet temper-

ature decreases significantly as expected.

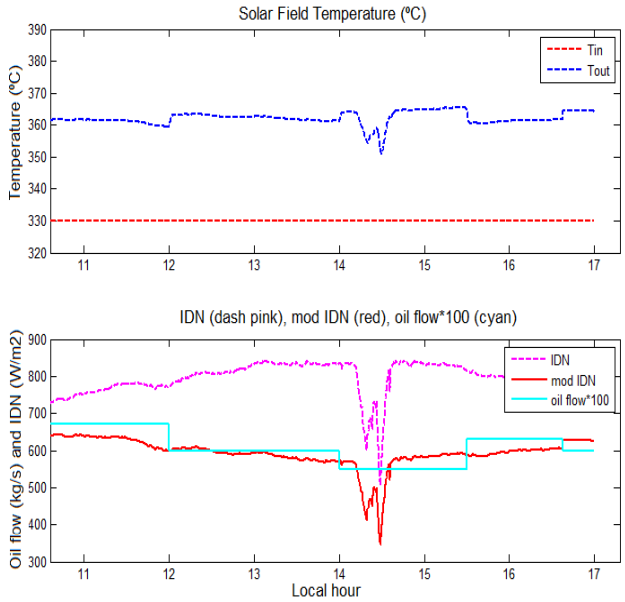


Figure 7. Simulation of a day with scattered clouds.

5 Conclusions

In order to develop efficient and robust control techniques, the use of accurate mathematical models of the palnt becomes necessary. A mathematical model of the solar field of the new TCP-100 research facility at the Plataforma Solar de Almería was presented. The simulations of the dynamic model under different operating conditions showed that the model behaved as expected.

As can be seen in figure 4, for the nominal conditions (IDN=900 W/m^2, T_{in}=330 and the mass flow=18.72 kg/s), the model produces a similar outlet temperature to that expected (380 °C).

When experimental data is available and the model adjustment is completed, this will be a very important testbench for simulating advanced control strategies and optimization algorithms.

Acknowledgment

The authors would like to thank to the CIEMAT Research Centre for supplying useful information to develop the model. We would also thank to the Junta de Andalucía and the European union for partially funding this work under the projects "Gestión óptima de edificios de energía cero" (P11-TEP-8129) and " Dynamic Management of Physically Coupled Systems of Systems" (DYMASOS) (FP7-ICT-ICT-2013.3.4-611281), EU 7^{th} Framework Programme (Theme Energy 2012.2.5.2) under grant agreement 308912 - HYSOL project - Innovative Configuration of a Fully Renewable Hybrid CSP Plant; and the National R+D+i Plan Project DPI2014-56364-C2-2-R of the Spanish Ministry of Economy and Competitiveness and ERDF funds.

References

Extended moving boundary model for two-phase flows, Prague. Czech Republic, jul 2005.

Hans D. Baerh. *Tratado Moderno de Termodinámica*. Springer-Verlag, first edition, 1965.

M. Berenguel. *to the Control of Distributed Solar Collectors*. PhD thesis, Universidad de Sevilla, 1996.

M. J. Blanco and L. R. Santigosa. *Advances in concentrating solar thermal research and Technology*. Elselvier, 2017. doi:978-0-08-100517-0.

J. Bonilla. *Modeling of Two-phase Flow Evaporators for Parabolic-Trough Solar Thermal Power Plants*. Phd thesis, University of Almería, 2013.

J Bonilla, L J Yebra, and S Dormido. A heuristic method to minimise the chattering problem in dynamic mathematical two-phase flow models. *Mathematical and Computer Modelling*, 54(5-6):1549–1560, 2011.

J Bonilla, L J Yebra, S Dormido, and E Zarza. Parabolic-trough solar thermal power plant simulation scheme, multi-objective genetic algorithm calibration and validation. *Solar Energy*, 86(1):531–540, 2012.

E. F. Camacho and M. Berenguel. Control of solar energy systems. In *8th IFAC Symposium on Advanced Control of Chemical Processes*, pages 848–855, 2012.

E. F. Camacho and A. J. Gallego. Optimal operation in solar trough plants: a case study. *Solar Energy*, 95:106–117, 2013.

E. F. Camacho and A. J. Gallego. Model predictive control in solar trough plants: A review. In *5th IFAC Conference on Nonlinear MPC, September 17-20*, Sevilla (Spain), 2015.

E. F. Camacho, F. R Rubio, and M. Berenguel. *Advanced control of solar plants*. Springer-Verlag, 1997.

E. F. Camacho, T. Samad, M. Garcia-Sanz, and I. Hiskens. Control for renewable energy and smart grids. Technical report, IEEE Control Systems Society, 2011.

E. F. Camacho, M. Berenguel, and A. J. Gallego. Control of thermal solar energy plants. *Journal of process control*, 2013.

Eduardo F. Camacho, M. Berenguel, Francisco.R. Rubio, and D. Martínez. *Control of Solar Energy Systems*. Springer-Verlag, 2012.

E.F. Camacho, F.R. Rubio, M. Berenguel, and L. Valenzuela. A survey on control schemes for distributed solar collector fields. part I: Modeling and basic control approaches. *Solar Energy*, 81:1240–1251, 2007.

Ricardo Carmona. *Análisis, Modelado y control de un campo de colectores solares distribuidos con sistema de seguimiento en un eje.* PhD thesis, Universidad de Sevilla, 1985.

J. Duffie and J. Beckman. *Solar engineering of thermal processes*. Wiley-Interscience, 2nd edition, 1991a.

A. J. Gallego and E. F. Camacho. Adaptative state-space model predictive control of a parabolic-trough field. *Control Engineering Practice*, 20 (9):904–911, 2012.

A. J. Gallego, F. Fele, E. F. Camacho, and L. J. Yebra. Observer-based model predictive control of a solar trough plant. *Solar Energy*, 97:426–435, 2013.

D. Yogi Goswami, F. Kreith, and J. F. Kreider. *Principles of Solar Engineering*. 2nd edition, 2000.

Ulf Herrmann and David W. Kearney. Survey of thermal energy storage for parabolic trough power plants. *Journal of Solar Energy Engineering*, 124:145–152, May 2002.

IRENA. Renewable energy technologies: Cost analysis series: Concentrating solar power. Technical report, International Renewable Energy Agency, 2012.

M.C. EL JAI and F.Z. CHALQI. A modified model for parabolic trough solar receiver. *American Journal of Engineering Research*, 02 (05):200–211, 2013.

J.M. Lemos, L.M. Rato, and E. Mosca. Integrating predictive and switching control: Basic concepts and an experimental case study. In F. Allgower and A. Zheng, editors, *Nonlinear Model Predictive Control*, number 1, pages 181–190. 2000.

S. D. Oden and H. I. Abu-Mulaweh. Design and development of an educational solar tracking parabolic trough collector system. *Global Journal of Engineering Education*, 15 (1):21–27, 2013.

R. Osterholm and J. Palsson. Dynamic modelling of a parabolic trough solar power plant. In *Proceedings of the 10th International Modelica Conference*, Lund, Sweden, March 10-12 2014.

F.R. Rubio, E. F. Camacho, and R. Carmona. Control de campos de colectores solares. *RIAI Revista Iberoamericana de Automática e Informática Industrial*, 3:26–45, 2006.

L.J. Yebra, M. Berenguel, E. Zarza, and S. Dormido. Object oriented modelling and simulation of parabolic trough collectors with modelica. In *5th MathMod Conference 2006*, volume 14, pages 361–375, Vienna University of Technology. Austria, feb 2006. Taylor & Francis. doi:10.1080/13873950701847199.

I. H. Yilmaz and M. S. Soylemez. Thermo-mathematical modeling of parabolic trough collector. *Energy Conversion and Management*, 88:768–784, 2014.

Mathematical Conditions in Heliostat Models for Deterministic Computation of Setpoints

Moisés Villegas-Vallecillos[1] Luis J. Yebra[2,3]

[1]Departamento de Matemáticas, Universidad de Cádiz, Spain, moises.villegas@uca.es
[2]Plataforma Solar de Almería, CIEMAT, Spain, luis.yebra@psa.es
[3]CIESOL, Joint Centre of the University of Almería-CIEMAT, Spain

Abstract

In this paper a set of mathematical conditions on heliostat models is presented. Its purpose is to guarantee a deterministic computation of the heliostat setpoints in azimuth (β) and elevation (α). In Central Receiver (CR) Concentrating Solar Power (CSP) plants, thousands of heliostats are continuously operated, and the updating of their setpoints is required frequently. For this reason, the fulfillment of some mathematical conditions of the mentioned type is important. In a simplified approach, during the operation, each heliostat reflects in its mirror a ray from the sun that impacts on a given aiming point P. This aiming point is assumed to be higher than the heliostat position, in the tower receiver. If v is the incident solar vector, x is the orthogonal vector of the heliostat reflective plane and $f(x)$ is the center of the heliostat mirror, then a system of equations with unknown x is arisen. Imposing certain conditions on f, we can ensure the existence and uniqueness of solution of this system, and provide a sequence converging to such solution. Furthermore, we offer a numerical method for approximating the solution in a deterministic form, which can be computed with the requirements of hard real time systems.

Keywords: central receiver concentrating solar thermal plants, heliostat setpoint, Banach's fixed point theorem, Newton-Raphson's numerical method, deterministic computation

1 Introduction

From a simplified point of view, a heliostat could be considered as a device that supports a mirror and orientates it, following the daily motion of the sun through a servomechanism and a multibody system (Otter et al., 2003). The mirror should be continuously properly oriented to reflect the sun radiation into an aiming point (objective point) positioned in the receiver of a Central Receiver (CR) Concentrating Solar Power (CSP) plant. In this way, all the heliostats in a CR CSP plant will concentrate the primary power from the sun into a set of predefined aiming points, that are carefully selected to generate a distributed power density pattern in the receiver. Figures 1 and 2 show this scenario.

Some authors have studied different ways to improve the focusing system of the heliostats. For example, Mulholland presents in (Mulholland, 1983) a method to determine the optimum size of the heliostat or concentrator, Guo et al. give in (Guo et al., 2012) a least squares model to determine certain angular errors in some altitude-azimuth tracking formulas, García-Martín et al. develop in (García-Martín et al., 1999) a control strategy aimed at optimizing the temperature distribution within a volumetric receiver, Berenguel et al. apply in (Berenguel et al.,

Figure 1. CESA-I central receiver research facility at Plataforma Solar de Almería (PSA).

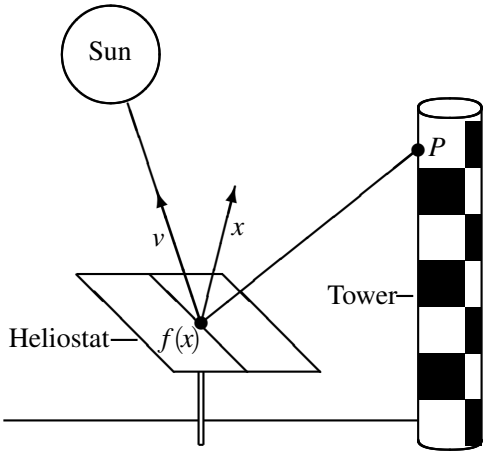


Figure 2. Simplified scheme showing the reflection of a ray representing the direct component of solar radiation being reflected in the heliostat mirror and impacting in one of the aiming points P of the receiver.

2004) a correction control system using artificial vision techniques, etc.

A related issue is to properly distribute the reception points of the tower where solar energy is focused. So it will be possible to get a uniform temperature profile in the receiver and prevent its structure from being damaged. To achieve that and develop a robust control system for the heliostat field, the following problem is considered:

Given a heliostat in which the center C and the orthogonal vector x of its mirror plane are variable, how can be determined them so that the reflected ray impacts on a specific receiver point P?

When the center C of the heliostat mirror can be expressed as a function $f(x)$ of the orthogonal vector x, it is possible to raise a system of equations to determine x (Section 2). In such case, some conditions are given on f to ensure the existence and uniqueness of solution of such system, and to obtain a sequence converging to that solution (Section 3). Concretely, it is required that f is a Lipschitz function and its Lipschitz constant verifies certain boundedness (see Condition 3.3).

Furthemore, a numerical method is developed to approximate the solution by mixing the aforementioned convergent sequence with Newton–Raphson's method. Thus, a safe and good speed convergence is obtained (Sections 4 and 5), making the numerical method a potential solution to use in hard real time control systems (Burns and Wellings, 2010).

2 Approach to the system of equations

In principle, it must be considered the relationship between the incident ray, the reflected ray and the reflective surface of the heliostat. That relationship is given by the law of reflection (Hecht, 2002, pages 98 and 99).

Law 2.1 (The Law of Reflection) *The incident ray, the normal to the reflective surface and the reflected ray all lie in the same plane. Furthermore, the angle of reflection is equal to the angle of incidence.*

This law and Proposition 2.3 are key tools to raise a system of equations. Before stating that proposition, a lemma is presented that will simplify its proof. Throughout all this paper, $(\cdot|\cdot)$ will denote the usual Euclidean inner product of $\mathbb{R}^N$ (where N is some positive integer), and $\|\cdot\|$ will represent the norm induced by this inner product. On the other hand, given $v_1,\dots,v_m \in \mathbb{R}^N$, we denote by $\mathrm{lin}(\{v_1,\dots,v_m\})$ the linear subspace spanned by $\{v_1,\dots,v_m\}$ (see (Larson and Falvo, 2012) for details).

Lemma 2.2 *Let $u,w \in \mathbb{R}^N$ vectors with $\|u\| = 1 = \|w\|$ and $u \neq -w$. If θ is the angle formed by the vectors u and w, then the angle between u and $u+w$ is $\theta/2$. As a consequence, $(u|u+w) \geq 0$.*

Next, the announced key proposition is presented.

Proposition 2.3 *Let $v, P, C \in \mathbb{R}^3$ such that v and $P-C$ are linearly independent, and let $x = P-C+(\|P-C\|/\|v\|)v$. It is satisfied that:*

i) If the heliostat reflective plane H passes through the point C and has orthogonal vector x, then the plane $C+\mathrm{lin}(\{v,x,P-C\})$ is perpendicular to H.

ii) The angle between $P-C$ and x is equal to the angle between v and x.

iii) The vectors v and $P-C$ are in the half-space determined by x, that is, $0 \leq (v|x)$ and $0 \leq (P-C|x)$.

If the incident solar ray passes through the point C and has direction vector v, and the heliostat reflective plane H passes through C and has orthogonal vector

$$x = P-C+(\|P-C\|/\|v\|)v, \tag{1}$$

then this proposition and the law of reflection state that the reflected ray passes through C and P.

It is not possible that v and $P-C$ are linearly dependent because the tower where P is situated prevents (with its shadow) the existence of an incident solar ray passing through P and C. But even if v and $P-C$ are linearly dependent and there exits the incident solar ray, the reflected ray and the line through C and P are coincident when H has orthogonal vector $x = P-C+(\|P-C\|/\|v\|)v$ (in fact, they would also coincide with the incident ray and the normal to H). Therefore, it will not be necessary to suppose that v and $P-C$ are linearly independent in subsequent developments.

Conditions 2.4 *From now on we assume the following conditions given by the nature of the problem:*

a) The director vector $v = (a_1,a_2,a_3)$ of the incident solar ray satisfies that $a_3 \geq 0$.

b) We suppose that the center C of the heliostat mirror can be expressed as a function of its orthogonal vector, that is, there exists a function $f\colon \mathbb{R}^3\backslash\{0\} \to \mathbb{R}^3$ such that, if $x \in \mathbb{R}^3\backslash\{0\}$ is a orthogonal vector to the heliostat mirror, then $C = f(x)$ is the center of such mirror. Note that, in such case, it holds $f(rx) = f(x)$ for all $x \in \mathbb{R}^3\backslash\{0\}$ and for all $r > 0$.

c) Whatever the center of the heliostat mirror, the distance from it to the impact point of the tower is always less than certain upper bound, that is, there is $R > 0$ such that $\|P-f(x)\| \leq R$ for all $x \in \mathbb{R}^3\backslash\{0\}$.

d) The impact point of the tower $P = (p_1,p_2,p_3)$ is quite higher than the heliostat, so there is a lower bound $M > 0$ such that $M \leq p_3 - f_3(x)$ for all $x \in \mathbb{R}^3\backslash\{0\}$.

Moreover, we consider the following subset of $\mathbb{R}^3$:

$$D_M = \{(x_1,x_2,x_3) \in \mathbb{R}^3 \colon M \leq x_3,\ \|(x_1,x_2,x_3)\| \leq 2R\}.$$

Taking into account Proposition 2.3 and the comments following it, if there exists $x \in \mathbb{R}^3 \backslash \{0\}$ such that

$$x = P - f(x) + \frac{\|P - f(x)\|}{\|v\|} v \tag{2}$$

and the heliostat reflective plane is orthogonal to x, then the ray reflected by the heliostat impacts on the point P. Therefore, we must solve the system of equations given by equality (2). Fortunately, we do not have to require too much of f to ensure the existence of solution of system (2). This is shown by the following result (which can be found in (Deimling, 1985, Theorem 3.2)).

Theorem 2.5 (Brouwer's Fixed Point Theorem) *Let D be a nonempty compact convex subset of $\mathbb{R}^N$ and let $g\colon D \to D$ be a continuous function. Then g has a fixed point, that is, there exists $x \in D$ with $x = g(x)$. The same is true if D is only homeomorphic to a compact convex set.*

Since $P - f(x) + (\|P - f(x)\|/\|v\|)\, v \in D_M$ for all $x \in D_M$, then Theorem 2.5 shows that system (2) has solution in D_M when it is assumed

Condition 2.6 *f is continuous on D_M.*

3 Sufficient conditions for uniqueness of solution and convergence

Before continuing with our development, we recall some concepts and an important result. Given $D \subseteq \mathbb{R}^N$ and a function $g\colon D \to \mathbb{R}^N$, it is said that g is Lipschitz on a subset $A \subseteq D$ if there is a real constant $c \geq 0$ such that

$$\|g(x) - g(y)\| \leq c\|x - y\| \tag{3}$$

for all $x, y \in A$. The least constant c for which the preceding inequality holds will be denoted by $L(g|_A)$. If $L(g|_A) < 1$, it is said that g is a contraction mapping on A.

A basic tool to prove existence and uniqueness of solutions of systems of equations is Banach's celebrated fixed-point theorem. It is true for general complete metric spaces but here we only need its version for closed subsets of $\mathbb{R}^N$ (see (Deimling, 1985, Theorem 7.1) or (Edwards, 1994, Theorem 3.1)).

Theorem 3.1 (Banach's Fixed Point Theorem) *Let D be a closed subset of $\mathbb{R}^N$ and $g\colon D \to D$ be a contraction mapping on D. Then:*

i) g has a unique fixed point $x \in D$.

ii) For all $x^{(0)} \in D$, the sequence $\left\{x^{(n)}\right\}$, given by

$$x^{(n+1)} = g\left(x^{(n)}\right) \qquad (n \in \mathbb{N} \cup \{0\}), \tag{4}$$

converges to x, and satisfy

$$\left\|x^{(n)} - x\right\| \leq \frac{L(g)^n}{1 - L(g)} \left\|x^{(1)} - x^{(0)}\right\| \quad (n \in \mathbb{N}). \tag{5}$$

Our aim in this section is to obtain a sequence converging to a solution of system (2) by using Theorem 3.1. In the following proposition we give sufficient conditions on f to get a contraction mapping and be able to apply Theorem 3.1.

Proposition 3.2 *Let $v = (a_1, a_2, a_3), P = (p_1, p_2, p_3) \in \mathbb{R}^3$ be with $a_3 \geq 0$, $M, R \in \mathbb{R}$ be bounds with $0 < M < R$ and $f\colon \mathbb{R}^3 \backslash \{0\} \to \mathbb{R}^3$ be a function such that $M \leq p_3 - f_3(x)$, $\|P - f(x)\| \leq R$ and $f(rx) = f(x)$ for all $x \in \mathbb{R}^3 \backslash \{0\}$ and for all $r > 0$. Suppose that there exists $\delta \in (0, M]$ such that f is Lipschitz on the set*

$$D_\delta = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \colon \delta \leq x_3,\ \|(x_1, x_2, x_3)\| \leq 2R\}$$

and $L(f|_{D_\delta}) < M/(2\delta)$. Then the function $g\colon D_M \to D_M$, defined by

$$g(x) = P - f(x) + \frac{\|P - f(x)\|}{\|v\|} v \qquad (x \in D_M), \tag{6}$$

is a contraction mapping on D_M and $L(g) \leq 2\delta L(f|_{D_\delta})/M$.

The new condition given by Proposition 3.2 is

Condition 3.3 *There exists $\delta \in (0, M]$ such that f is Lipschitz on D_δ and $L(f|_{D_\delta}) < M/(2\delta)$.*

If we assume this condition, Theorem 3.1 and Proposition 3.2 tell us that system (2) has a unique solution in D_M, and provide us with a sequence converging to that solution.

Remark 3.4 *Supposing that the function f is Lipschitz on D_δ is not too restrictive. For example, if we assume that f is of class $\mathscr{C}^1$ on some open set containing D_δ, then f is Lipschitz on D_δ. We give a bound for the Lipschitz constant of f in this situation. As the partial derivatives of f are bounded on D_δ, we can find $k_1(\delta), k_2(\delta), k_3(\delta) \in \mathbb{R}_0^+$ such that*

$$\|\nabla f_j(x)\| \leq k_j(\delta) \qquad (x \in D_\delta,\ j = 1, 2, 3). \tag{7}$$

It can be proved by applying the Mean Value Theorem (Edwards, 1994, Theorem 3.4) that the Condition 3.3 is fulfilled if

$$k_1(\delta) + k_2(\delta) + k_3(\delta) \leq M/(2\delta). \tag{8}$$

4 Newton–Raphson's method for the raised system

In this section we recall Newton–Raphson's method for solving system (2). Suppose we want to solve a system of three equations with three unknowns of the form $F(x) = 0$, where F is a function from a subset $\Omega \subseteq \mathbb{R}^3$ into $\mathbb{R}^3$. One idea is to find a matrix $Q = (q_{ij})$ whose entries q_{ij} are functions from Ω into $\mathbb{R}$, such that the fixed-point iteration determined by the function $G\colon \Omega \to \mathbb{R}^3$, where

$$G(x) = x - Q(x)F(x) \qquad (x \in \Omega); \tag{9}$$

gives quadratic convergence to the solution of $F(x)=0$. The following theorem motivates the choice of Q (see (Burden and Faires, 2010, Theorem 10.7)). Its proof is based on the Taylor development of G (consult (Edwards, 1994, Theorem 7.1)).

Theorem 4.1 *Let Ω be an open subset of $\mathbb{R}^N$, $A \subseteq \Omega$ be a convex subset and $G:\Omega \to \mathbb{R}^N$ be a function of class $\mathscr{C}^2$ on Ω. Suppose that there exists a fixed point $w \in A$ of G, and G has the following properties:*

a) $\left|\partial^2 G_i(x)/(\partial x_j \partial x_k)\right| \leq K$ *for each* $x \in A$, $i,j,k \in \{1,\ldots,N\}$, *and for some constant* $K>0$ *(that is, the second partial derivatives of G are bounded on A).*

b) $\partial G_i(w)/\partial x_j = 0$ *for all* $i,j \in \{1,\ldots,N\}$.

Then, for all $x^{(0)} \in A$ with $\|x^{(0)}-w\| < 2/(N^{5/2}K)$, the sequence generated by

$$x^{(n+1)} = G\left(x^{(n)}\right) \qquad (n \in \mathbb{N}\cup\{0\}) \tag{10}$$

converges quadratically to w. Concretely,

$$\|x^{(n+1)}-w\| \leq \frac{N^{5/2}K}{2}\|x^{(n)}-w\|^2 \quad (n \in \mathbb{N}\cup\{0\}). \tag{11}$$

When $G(x) = x - Q(x)F(x)$, and Q,F are of class $\mathscr{C}^1$, the second condition of Theorem 4.1 is equivalent to $Q(w)J_F(w) = I_3$, where $J_F(w)$ is the Jacobian matrix of F at the point w and I_3 is the identity matrix. So Newton–Raphson's method consists in choosing $Q(x) = J_F(x)^{-1}$ for each $x \in \Omega$, and performing the fixed point iteration determined by the function G, where

$$G(x) = x - J_F(x)^{-1}F(x) \qquad (x \in \Omega). \tag{12}$$

Remark 4.2 *To apply Newton–Raphson's method and ensure its quadratic convergence to the solution of the system (using Theorem 4.1), we must impose the following conditions:*

i) *Ω must be an open neighbourhood of the solution w of the system $F(x)=0$.*

ii) *F has to be of class $\mathscr{C}^3$ on Ω (take in mind the inverse mapping theorem).*

iii) $\det(J_F(x)) \neq 0$ *for all* $x \in \Omega$.

iv) *The second partial derivatives of the function G must be bounded on a convex subset $A \subseteq \Omega$ containing w.*

v) *The first value of the iteration $x^{(0)} \in A$ has to be close enough to the solution w.*

Next we see how are these requirements in our context. Let $f\colon \mathbb{R}^3\backslash\{0\} \to \mathbb{R}^3$, $v,P \in \mathbb{R}^3$ and $R,M \in \mathbb{R}$ be as in Conditions 2.4, and let

$$D_M = \{(x_1,x_2,x_3) \in \mathbb{R}^3 \colon M \leq x_3,\ \|(x_1,x_2,x_3)\| \leq 2R\}.$$

Consider the function $F : \mathbb{R}^3\backslash\{0\} \to \mathbb{R}^3$ defined by

$$F(x) = P - f(x) + \frac{\|P-f(x)\|}{\|v\|}v - x \quad (x \in \mathbb{R}^3\backslash\{0\}). \tag{13}$$

Conditions 4.3 *The requisites above can be rewritten for system (2) as follows:*

a) *f must be of class $\mathscr{C}^3$ on an open subset $\Omega \subseteq \mathbb{R}^3\backslash\{0\}$ containing D_M.*

b) $\det(J_F(x)) \neq 0$ *for all* $x \in \Omega$.

c) *To perform the fixed point iteration given by the function $G\colon \Omega \to \mathbb{R}^3$, where $G(x) = x - J_F(x)^{-1}F(x)$ for all $x \in \Omega$, the first value $x^{(0)} \in D_M$ has to be close enough to the solution.*

Note that under these requirements, as D_M is compact and G is of class $\mathscr{C}^2$, it holds that the second partial derivatives of G are bounded on D_M.

5 A numerical method to approximate the solution

Under the conditions of Proposition 3.2, the development carried out in Section 3 gives us an iterative method to approximate the solution of system (2). This is the fixed point iteration given by the contraction mapping g. The convergence of this method is safe. However, the method converges with linear speed, whence it is slower than Newton–Raphson's method.

On the other hand, when f is of class $\mathscr{C}^3$ on an open subset containing D_M, we have no guarantees that Newton–Raphson's method can be applied because, a priori, we do not know if $\det(J_F(x)) \neq 0$. Moreover the convergence of Newton–Raphson's method is local, since $x^{(0)}$ must be chosen close enough to the solution.

For these reasons, we try to use both methods together in order to take advantage of both. First, we get close to the solution using the contraction mapping method. We must set an initial tolerance $tol0$, and, when it is attained, we change over to Newton–Raphson's method. If Newton–Raphson's method can not be applied, we continue with the contraction mapping method. In Algorithm 1 we can see the procedure.

Note that, to apply this algorithm, the function f of the heliostat and Newton–Raphson's method must first be implemented. An outline of Newton–Raphson's method can be found in (Burden and Faires, 2010, Algorithm 10.1). To do its implementation, take into account that the method fails when the Jacobian matrix of F is not regular, there is no convergence ($2tol0 \leq \left\|x^{(k+1)}-x^{(k)}\right\|$) or the maximum number of iterations is exceeded.

Algorithm 1 Numerical method to approximate the solution

Input: v, P, M, δ, Lf, $x0$, $tol0$, tol, m.
Output: w, $iter0$, $iter1$, $iter2$.

1: $Lg \leftarrow 2\,\delta\,Lf/M$
2: $xk \leftarrow x0$
3: $w \leftarrow P - f(xk) + (\|P - f(xk)\|/\|v\|)v$ ▷ second value of the iteration
4: **if** $\|w - xk\| = 0$ **then**
5: $\quad iter0 \leftarrow 1$
6: **else** ▷ calculate the number of iterations to attain $tol0$:
7:

$$iter0 \leftarrow 1 + \text{floor}\left(\log\left(\frac{(1-Lg)\,tol0}{\|w-xk\|}\right)/\log(Lg)\right)$$

8: **end if**
9: $k \leftarrow 1$
10: **while** $k < iter0$ and $k < m$ and $\|w - xk\| \neq 0$ **do**
11: $\quad xk \leftarrow w$
12: $\quad w \leftarrow P - f(xk) + (\|P - f(xk)\|/\|v\|)v$
13: $\quad k \leftarrow k + 1$
14: **end while**
15: $iter0 \leftarrow k$
16: Apply Newton–Raphson's method with:

Input: $w, tol0, tol, m, v, P$.
Output: w, $iter1$, control variable to indicate if Newton–Raphson's method fails.

17: **if** Newton–Raphson's method fails **then**
18: $\quad xk \leftarrow w$
19: $\quad$ do Steps 3–15 with tol and $iter2$ instead of $tol0$ and $iter0$
20: **else**
21: $\quad iter2 \leftarrow 0$
22: **end if**

Output: $w, iter0, iter1, iter2$.

6 First example. Type B heliostats

In type B heliostats there is a unique joint between two pieces: the post and the arm (see Fig. 3). The heliostat mirror is at the end of the arm so that the mirror plane is perpendicular to the arm. Let Q be the point where the joint is situated and l be the length of the arm. If x is a vector orthogonal to the mirror plane, then the center of the heliostat mirror is given by

$$f(x) = Q + \frac{l}{\|x\|}\,x \tag{14}$$

As f is continuous, Theorem 2.5 ensures the existence of solution of system (2). In fact, given $\delta \in (0, M]$, f is Lipschitz on D_δ and $L\left(f|_{D_\delta}\right) \leq 2l/\delta$. Then Proposition 3.2 can be applied if it is satisfied that $2l/\delta < M/(2\delta)$ for some $\delta \in (0, M]$. Taking, for example, $\delta = M$, this inequality is fulfilled since $4l < M$. Therefore Proposition 3.2 and Theorem 3.1 give us existence and uniqueness

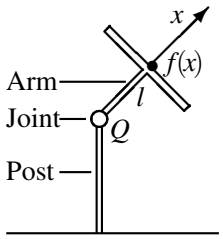


Figure 3. Type B heliostat

of solution of the system (2) on D_M, and a fix point iteration converging to it. Moreover $L(f|_{D_M}) \leq 2l/M$ and $L(g) \leq 4l/M$.

Besides, to apply Newton–Raphson's method, we need the partial derivatives of the function $F\colon \mathbb{R}^3\backslash\{0\} \to \mathbb{R}^3$ given, for each $x \in \mathbb{R}^3\backslash\{0\}$, by

$$F(x) = P - Q - \frac{l}{\|x\|}\,x + \left\|P - Q - \frac{l}{\|x\|}\,x\right\|\frac{v}{\|v\|} - x. \tag{15}$$

We prove straightforwardly that the Jacobian matrix of F is

$$J_F(x) = \frac{l}{\|x\|}\left(\frac{xx^T}{\|x\|^2} - I_3\right) + \frac{l\,v\left(P - Q - \frac{l}{\|x\|}x\right)^T}{\|x\|\|v\|\left\|P - Q - \frac{l}{\|x\|}x\right\|}\left(\frac{xx^T}{\|x\|^2} - I_3\right) - I_3 \tag{16}$$

where the vectors are considered as 3×1 column matrices and I_3 is the identity matrix.

Example 6.1 Consider two type B heliostats, one with its joint at the point $Q_1 = (-0.008, 186.996, 5.445)$ and other with its joint at $Q_2 = (-35.191, 383.648, 11.125)$. Both heliostats have an arm of a meter in length. Two solar rays can strike on them. Their direction vectors are $v_1 = (0.591431, -0.164578, 0.789382)$ and $v_2 = (0.510816, -0.199895, 0.836127)$. It is desired that, for both solar incident rays, the reflected rays from each heliostat impact first on the point $P_1 = (20, 1.7, 75)$, after on the point $P_2 = (-0.6, 1.738, 73.991)$ and finally on $P_3 = (0.008, 6.524, 34.165)$ (all coordinates are in meters). We can take the bound $M = 34 - 13 = 21$. The obtained results are shown in Table 1. Remember that our numerical method (Algorithm 1) consists of three stages. In Table 1, the number of iterations of the first stage (fixed point iteration of the function g) is represented by FP, for the second stage (Newton-Raphson's method) we put NR, and for the last stage (again fixed point iteration of the function g) FP is used too.

Finally, we compare the hybrid method and Newton-Raphson's method. Taking the first heliostat, the incident solar vector v_1 and the impact point P_1, then the solution $x = (136.7502, -217.2572, 225.2455)$ is obtained

Table 1. Results with type B heliostats

x_1	x_2	x_3	FP	NR	FP
136.7502	−217.2572	225.2455	4	2	0
115.3882	−216.9750	223.1364	4	2	0
107.2535	−209.6961	171.7516	4	2	0
120.8285	−224.2384	234.4766	4	2	0
99.5775	−223.9091	232.3027	4	2	0
92.6350	−216.1090	180.2265	4	2	0
285.6065	−445.4998	371.4253	5	1	0
263.5169	−445.0280	368.3868	5	1	0
258.7250	−438.7159	321.4195	4	2	0
254.1878	−459.2700	389.6425	5	1	0
232.3052	−458.7087	386.4839	5	1	0
228.2491	−452.0741	339.0899	4	2	0

Table 2. Comparison of Hybrid and Newton-Raphson's methods

	$iter0$	$iter1$	$iter2$
Hybrid	4	2	0
N.R.		5	

with both methods. However the number of iterations employed by each method changes. This is shown in Table 2. The columns $iter0$, $iter1$ and $iter2$ represent the number of iterations in each of the stages of the hybrid method as detailed in Algorithm 1. For Newton-Raphson's method, $iter1$ correspond to the total number of effected iterations.

The iterations of Newton-Raphson's method are fewer, but we must observe that:

1. Once the first iterations of the hybrid method are passed, it is also used Newton–Raphson's method.

2. At each step of Newton–Raphson's method a Jacobian matrix must be calculate and a linear system must be solved. Hence the number of operations of Newton-Raphson's method is greater than the number of operations of the fixed point iteration given by the contraction mapping g.

3. There is no guarantees that Newton–Raphson's method can be always applied.

7 Second example. Type A heliostats

In this section, we consider the simplest heliostat type. Now the center of the mirror is fixed, that is, there is $C=(c_1,c_2,c_3)\in\mathbb{R}^3$ such that $f(x)=C$ for all $x\in\mathbb{R}^3\backslash\{0\}$. Then system (2) is trivially solved without applying any numerical method because we have

$$x = P - C + \frac{\|P-C\|}{\|v\|}\, v\,. \tag{17}$$

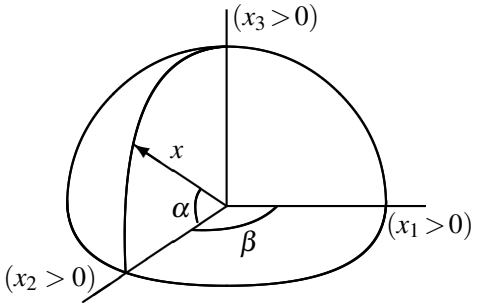


Figure 4. Spherical coordinates

It may be that a spherical coordinates of the orthogonal vector x are needed. For example, the angles $\alpha\in[-\pi/2,\pi/2]$ and $\beta\in(-\pi,\pi]$ determined by Fig. 4 may be required. In the following example these angles are calculated.

Example 7.1 A simple heliostat has the center of its mirror at the point $C_1=(-0.008,186.996,5.445)$, and another has it at the point $C_2=(-35.191,383.648,11.125)$. Two solar rays can strike on them, one with direction vector $v_1=(0.591431,-0.164578,0.789382)$ and other with direction vector $v_2=(0.510816,-0.199895,0.836127)$. It is desired that, for both solar incident rays, the reflected rays from each heliostat impact on the points $P_1=(20,1.7,75)$, $P_2=(-0.6,1.738,73.991)$ and $P_3=(0.008,6.524,34.165)$ (all coordinates are in meters). The implemented computations give the results presented in Table 3.

8 Conclusions

If a heliostat meets that the center C of its mirror can be expressed as a function $f(x)$ of its orthogonal vector x, and such function f satisfies certain general conditions (condition 3.2), then, given an incident solar vector v and a specific receiver point P, there exists x making the reflected ray impacts on P. Furthermore, a numerical method is developed to approximate the solution x, with safe and good speed convergence properties. These properties make the numerical method a potential solution to use in hard real time control systems for the deterministic computation of the heliostat setpoints.

As we do not work with a specific function f, the model proposed to find the orthogonal vector x can be applied to different types of heliostats. For each heliostat type a mathematical model should be derived to obtain the

Table 3. Results with type A heliostats

Center of heliostat mirror	Incident solar vector	Impact point of the tower	Angle α (radians)	Angle β (radians)
C_1	v_1	P_1	0.72094	−1.0076
C_1	v_1	P_2	0.73797	−1.0805
C_1	v_1	P_3	0.63113	−1.0965
C_1	v_2	P_1	0.74487	−1.0753
C_1	v_2	P_2	0.75955	−1.1510
C_1	v_2	P_3	0.65516	−1.1645
C_2	v_1	P_1	0.61247	−1.0001
C_2	v_1	P_2	0.61942	−1.0355
C_2	v_1	P_3	0.56352	−1.0373
C_2	v_2	P_1	0.63908	−1.0647
C_2	v_2	P_2	0.64510	−1.1014
C_2	v_2	P_3	0.59063	−1.1027

function f.

Acknowledgment

The authors gratefully acknowledge the funding support from CIEMAT Research Centre, EU 7^{th} Framework Programme (Theme Energy 2012.2.5.2) under grant agreement 308912 - HYSOL project - Innovative Configuration of a Fully Renewable Hybrid CSP Plant, the National R+D+i Plan Project DPI2014-56364-C2-2-R of the Spanish Ministry of Economy and Competitiveness and ERDF funds.

Basic symbols

$\|\cdot\|$: usual Euclidean norm of $\mathbb{R}^N$.
v: incident solar vector.
P: aiming point in the tower receiver.
C: center of the heliostat mirror.
x: orthogonal vector of the heliostat mirror plane.
f: function relating the orthogonal vector x and the center C of the heliostat mirror.
R: maximum distance between P and the center of the heliostat mirror (or greater).
M: minimum difference between the height of P and the height of the heliostat mirror center (or less).
δ: a positive constant in the interval $(0,M]$.
D_δ: set of vectors in the closed ball of center 0 and radius $2R$ whose third coordinate is greater than or equal to δ.
$L(g|_A)$: Lipschitz constant of the function g on the set A (when A is the domain of the function g, we put $L(g)$).
g: function given by $g(x)=P-f(x)+(\|P-f(x)\|/\|v\|)v$.
∇f_j: gradient of the function f_j.

References

M. Berenguel, F. R. Rubio, A. Valverde, P. J. Lara, M. R. Arahal, E. F. Camacho, and M. López. An artificial vision-based control system for automatic heliostat positioning offset correction in a central receiver solar power plant. *Solar Energy*, 76: 563–575, 2004.

R. L. Burden and J. D. Faires. *Numerical analysis*. Brooks Cole, Boston, 9th edition, 2010.

A. Burns and A. J. Wellings. *Real-time systems and programming languages*, volume 2097. Addison-Wesley, 2010.

K. Deimling. *Nonlinear Functional Analysis*. Springer-Verlag, Berlin, 1985.

C. H. Edwards. *Advanced Calculus of Several Variables*. Dover books on advanced mathematics. Dover Publications, revised edition, 1994.

F. J. García-Martín, M. Berenguel, A. Valverde, and E. F. Camacho. Heuristic knowledge-based heliostat field control for the optimization of the temperature distribution in a volumetric receiver. *Solar Energy*, 66 (5): 355–369, 1999.

M. Guo, Z. Wang, J. Zhang, F. Sun, and X. Zhang. Determination of the angular parameters in the general altitude-azimuth tracking angle formulas for a heliostat with a mirror-pivot offset based on experimental tracking data. *Solar Energy*, 86 (3): 941–950, 2012.

E. Hecht. *Optics*. Addison-Wesley, New York, 4th edition, 2002.

R. Larson and D. Falvo. *Elementary Linear Algebra*. Belmont, CA: Brooks/Cole Cengage Learning, 7th international edition, 2012.

G. P. Mulholland. Determination of heliostat and concentrator size for solar furnace facilities. *Journal of Solar Energy Engineering*, 105 (3): 243–245, 1983.

M. Otter, H. Elmqvist, and S. E. Mattsson. The new modelica multibody library. In *Proceedings of the 3rd International Modelica Conference*. Citeseer, 2003.

Object-Oriented Dynamic Modelling of Gas Turbines for CSP Hybridisation

Luis J. Yebra[1,7] Sebastián Dormido[2] Luis E. Díez[3] Alberto R. Rocha[4] Lucía González[4]
Eduardo Cerrajero[5] Silvia Padilla[6]

[1]Plataforma Solar de Almería, CIEMAT, Spain, luis.yebra@psa.es
[2]Dpto. Informática y Automática, UNED, Spain, sdormido@dia.uned.es
[3]SERLED Consultores, Spain, le.diez@serled.com
[4]ACS-Cobra T&I, Spain, {arocha@acsindustria.com , luciagonzalez@grupocobra.com}
[5]IDIE (Investigación, Desarrollo e Innovación Energética) , Spain, eduardo.cerrajero@idie.es
[6]AITESA (Air Industrie Thermique) , Spain, spadilla@aitesa.es
[7]CIESOL, Joint Centre of the University of Almería-CIEMAT, Spain

Abstract

This paper presents a dynamic model of a gas turbine developed for the HYSOL project. The model is developed mainly for control purposes and based on mathematical, physical and chemical principles. Approximations and assumptions are presented with the objective to minimize complexity and to maintain a modular structure. The main modules are presented independently and ready to be connected to form the complete and parameterizable gas turbine model. Possible cases of algebraic loops appearance are detected and solutions are proposed to avoid them. Moreover, first principles compression and expansion maps are developed to avoid non-linear algebraic loops. The Modelica modelling language and the libraries Modelica.Fluid and Modelica.Media have been extensively used for the models development. Results from simulation experiments are presented, implementing the proposed mathematical models for compressor and turbine submodules independently, as well as for the complete gas turbine system.

Keywords: CSP hybridisation, gas turbine, object oriented modelling, first principle compression/expansion maps, simulation of gas turbines

1 Introduction

Mathematical approximations applied in the dynamic modelling of a gas turbine for model based control purposes are presented in this paper. The hypothesis presented come from the first principles used in a wide range operation model to be used in HYSOL european project. The information presented in this paper comes exclusively from published sources and detailed through references mainly focused in the technology behind the Modelica modelling language. Suggested reads for the Modelica language are (Cellier, 1991; Åström et al., 1998; Fritzson, 2004). For turbines concepts it has been mainly used (Bathie, 1996; Kehlhofer et al., 2009; Gülen and Kim, 2014). An important reference of previous implementations concepts is the ThemoPower Modelica library (Casella and Leva, 2006).

In this paper, the modelling principles for the three main gas turbine subsystems are presented, oriented to be implemented in the Modelica modelling language. Modelica is a general purpose acausal object-oriented modelling language for physical systems modelling (Fritzson, 2004), and its evolution is extensively described in (Åström et al., 1998). The Modelica Standard Library (MSL) is frequently referenced for base classes and final use models to be directly reused in gas turbine subsystems components. `Modelica.Media` and `Modelica.Fluid` are the two most used packages from MSL in the presented work. For modelling and simulation works the Dymola© tool (3DS, 2016) has been used.

2 Description of a Basic Gas Turbine Model

This section presents the approximations used in the modelling the different parts, all obtained from public references. The two main phenomena background come from thermofluid and mechanical disciplines. Object oriented thermofluid modelling is described in (Tummescheit, 2002), which indeed has its base in the Computational Fluid Dynamics (CFD) methods detailed in (Patankar, 1980) and (Versteeg and Malalasekera, 1995). Currently the `Modelica.Fluid` library implement these concepts, see (Elmqvist et al., 2003). With respect to mechanical components it is only required to model 1-dimensional rotational mechanical systems and therefore the `Modelica.Mechanics.Rotational` package has been directly used.

Following the above references, the modelling methodology for thermofluid parts use two main concepts detailed in (Tummescheit, 2002):

- Control Volumes (CV). Stating mass and energy conservation, computing effort variables from flows.

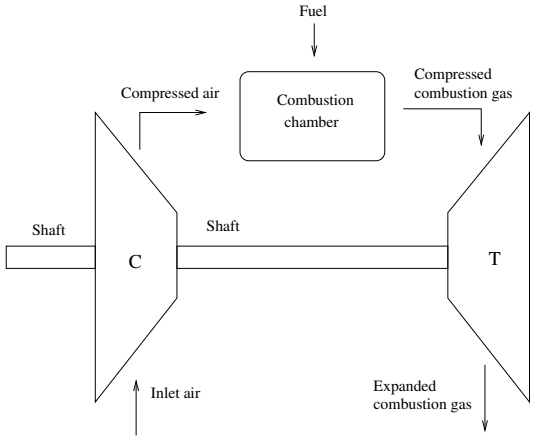


Figure 1. A general gas turbine diagram composed by compressor ('C'), combustion chamber, turbine ('T') and shaft. These are the minimum components to be considered for a basic turbine in an object oriented approach.

- Flow Models (FM). Stating momentum conservation, and computing flows from effort variables.

The general scheme regarded for the gas turbine is depicted in Fig. 1, referenced as basic gas turbine engine (Bathie, 1996), in which three main connected components are shown: compressor, combustion chamber and turbine. The compressor and turbine are connected mechanically by a shaft with other secondary components that are usually present in detailed models but not regarded in the analysis presented in this article.

For definition of effort and flow variables, see (Cellier, 1991). So all components will be classified as CV or FM, or an alternating arrangement of them: ...-CV-FM-CV-FM...

So all the components in this model will be classified in either of both categories: CV or FM, attending to minimize the order of the final differential and algebraic equation (DAE) system, that is, trying to minimize the number of components of type CV.

As detailed in (Kehlhofer et al., 2009) and (Bathie, 1996), the boundary conditions defined by environment air pressure, temperature and humidity strongly influence the behaviour of the turbine-set. Hereafter turbine-set is referred as the set composed by compressor-combustion chamber-turbine, that is our basic gas turbine. Interesting qualitative analysis about changes in operational point of the turbine-set w.r.t. environment variables changes is presented in two chapters in (Kehlhofer et al., 2009). So, any model to be used in a wide range conditions should accept variations at inlet air conditions. These variations should be compensated by the control system. The `Modelica.Media` library (Casella et al., 2006) has been used to define the set of classes that implement different mediums properties.

The turbine-set model will be used in model based control applications that would require mathematical model inversion. This usage obligates to apply assumptions that minimize model complexity and let to obtain the required mathematical inverse model.

2.1 Compressor

There exist different types of compressors with a variety of complexities and applications: axial, rotary, centrifugal and reciprocating. The model presented in this section is general and simpler than (Greitzer, 1976), assuming that the compressor is an idealized FM device in which neither mass nor energy conservation are stated, and the isentropic conditions are met in the operating range. Extreme situations like surge and stall instabilities, and choked flow conditions, are never met. Ideal gas state assumption is used for the air.

Isentropic conditions assume that the state of the gas changes through the compressor under the condition $ds = 0$ for any general formulation of balances of mass and energy. The general definition for specific entropy s appears as its classical definition equation:

$$Tds = du + pdv \tag{1}$$

where, in a differential mass element dm, T is the temperature, u is the specific internal energy, p the pressure, $v = 1/\rho$ the specific volume and ρ the density.

From the mechanical point of view, inside the compressor, there is an interaction between the gas flow (assumed isentropic flow) and the shaft blades that is stated in terms of a steady state energy balance as shown in (2):

$$\dot{m}\Delta h = \tau\omega\eta_{mec} \tag{2}$$

where $\dot{m}$ is mass flow rate, $\Delta h = h_{outlet} - h_{inlet}$ is the specific enthalpy increment through the compressor, τ is the torque applied over the shaft, ω the angular velocity of the shaft and η_{mec} the mechanic efficiency. The previous equation is actually an energy balance not usual in a FM, but necessary to *link* the thermodynamic and mechanical domains.

In practice, the parameters available from industrial compressors are usually implicit in the measured compressor characteristic (or map) from the company manufacturer. The compressor map is a 2-D table with its steady state operation points, that relates the pressure ratio Φ_c with the mass flow rate $\dot{m}$ and angular velocity ω. From the map, the $\Phi_c(\dot{m},\omega)$ could be fitted from a polynomial interpolation in these two *supposed independent* variables $\dot{m}$ and ω. So the next equation could be assumed as *constitutive* for the component:

$$\frac{p_o}{p_i} = \Phi_c(\dot{m},\omega) \tag{3}$$

where p_o and p_i are the output and input pressures of the compressor.

In the case of the computation of the specific enthalpy increment there are several approximations in the literature. Under isentropic conditions (4) is used:

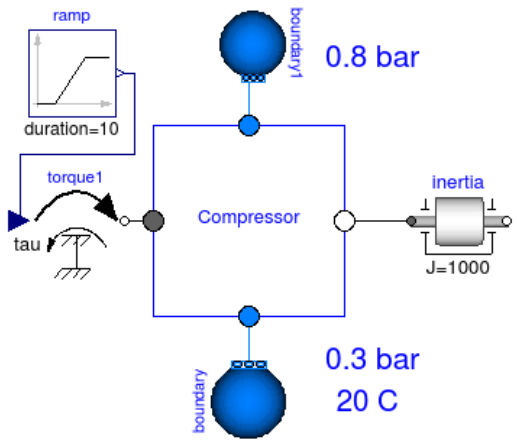


Figure 2. Diagram of the compressor in Dymola tool. The model implements the compressor map from one example of ThermoPower library.

$$\Delta h \approx \Delta h_s = \left(\left(\frac{p_o}{p_i} \right)^{\frac{\kappa-1}{\kappa}} - 1 \right) c_p T_i \tag{4}$$

where Δh_s is the enthalpy increment under isentropic conditions, κ the isentropic exponent, c_p is the specific heat at constant pressure, and T_i gas inlet temperature.

In those cases in which the approximation $\Delta h \approx \Delta h_s$ is not acceptable (5) could be applied, where η is the isentropic efficiency.

$$\eta(\dot{m}, \omega) = \frac{\Delta h_s}{\Delta h} \tag{5}$$

Isentropic efficiency is a part of the compressor maps too, depends on $(\dot{m}, \omega)$, and can be considered as another *constitutive* equation for the component.

In general, compressor maps are formed by $\Phi_c(\dot{m}, \omega)$, $\eta(\dot{m}, \omega)$, among other data. These maps are obtained by experimentation in steady state conditions assuming that $\dot{m}$ and ω are independent variables in the test. This experimentation produces data sets that not always perform the best under dynamic simulations and transformation to other independent variables set is recommended. Transformation to beta lines independent variables is usually applied for turbine specific simulation software.

Fig. 2 presents a schematic diagram in Dymola showing the experiment simulated for the compressor. The compressor model is implemented using isentropic approximations and isentropic efficiency, using all components from `Modelica.Fluid` and `Modelica.Media`. The gas used as medium is dry air as ideal gas.

Fig. 3 shows the evolution of angular velocity of the shaft `omega`, and the mass flow rate `mdot_compressor` are the simulated variables, under the shown boundary conditions in which the driving torque `Tu_ext` is the input. The compressor is loaded by an inertia component from `Modelica.Mechanics.Rotational` representing the shaft whose initial angular velocity is computed by Dymola. The torque positive ramp increments the angular velocity under constant thermofluid boundary conditions.

2.2 Combustion Chamber

This component performs the combustion of pressurized air with the fuel used. The usual fuel used in gas turbines are hydrocarbons, with a general chemical formula of C_xH_y. These reactions generate heat flow incoming from enthalpy of combustion.

Dynamic modelling of chemical reactions in general is clearly exposed in (Cellier, 1991), in the former modelling language for the Dymola tool. The dominant time constants of these kind of reactions are considerably smaller than the lowest one from thermofluid and mechanic dynamics. When the chemical reaction models are regarded, the DAE finally obtained for the complete gas turbine model become too stiff. The stiffness is a numerical property usually not welcome, but has to be accepted when it is needed to know internal information of the reaction evolution. Due to the main objective of this model is to predict thermal and mechanic dynamics, it was preferred to neglect the chemical dynamical behaviour, considering it as a non modelled dynamics. So, in this case, no composition changes are assumed in this model.

In this way, using Modelica components, the model used for the combustion chamber was a CV from `Modelica.Fluid(.Vessels.ClosedVolume)`, in which a prescribed heat flow is injected, emulating the heat flow incoming from continuous hydrocarbon burning gas flow. The heat flow will be calibrated with experimental data from the facility, depending on the hydrocarbon composition and flow rate. Fig. 4 shows the composition diagram.

Further modelling work for this component will be reaction modelling for the prediction of the composition variation, when needed.

2.3 Turbine Module

There are two general types of turbines, the radial-flow and axial-flow turbines, as detailed in (Bathie, 1996). In industrial applications the turbine module is composed of several submodules connected. But beyond the technical and engineering details behind turbines, the interesting point is that it is a device with deep similarities in the physical behaviour with the compressor. Although in this case, the gas is expanded when passing through it making mechanical work on the shaft.

For the previous reasons the turbine module is considered a FM in thermofluid modeling methodology, although again, an internal steady state energy balance must be formulated to *link* thermofluid and mechanical domains. So, (2) applies for both: turbine and compressor.

In general, from first principles point of view, the mathematical model presented in section 2.1 is applied in the turbine module, with the same requirement that the tur-

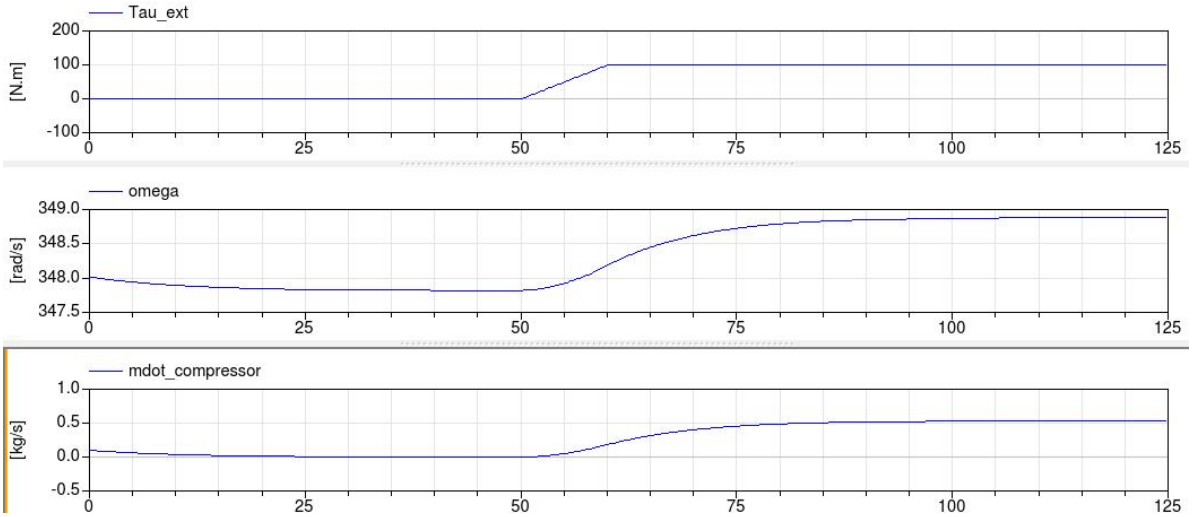


Figure 3. Simulation of the compressor under constant thermofluid boundary conditions and a ramp in the driving torque.

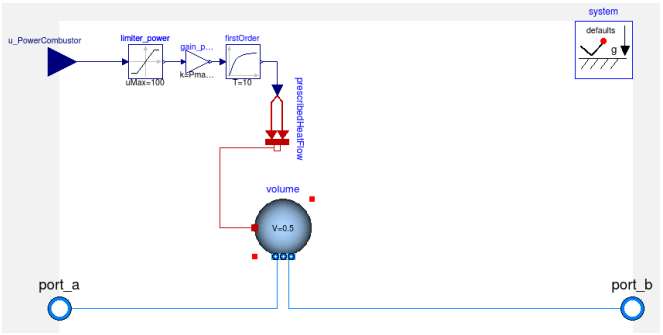


Figure 4. Diagram of the combustor in Dymola tool. The model implements a CV in which the air is heated by the combustion enthalpy, estimated outside the component.

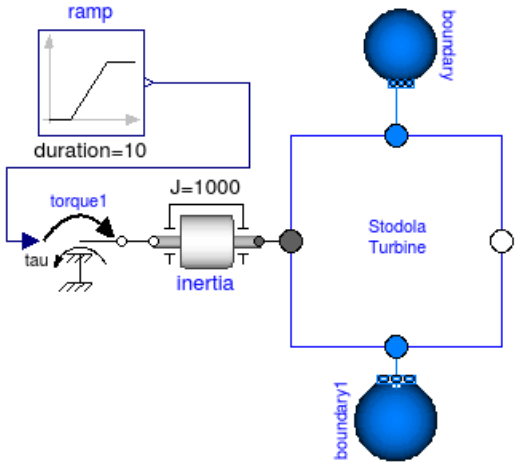


Figure 5. Diagram of the turbine in Dymola tool. The model implement the turbine map from one example of ThermoPower library.

bine *maps* $(\Phi_c(\dot{m},\omega),\eta(\dot{m},\omega))$ must be provided. These maps are obtained usually from steady state measurements under certain conditions (although in the turbine case the word *map* is not used so frequently as in the compressor case). So, from an object oriented point of view, could be possible to define the physical model once for both, the compressor and turbine module, in a partial base class from which both could be inherited and parameterized by their compression/expansion maps $(\Phi_c(\dot{m},\omega),\eta(\dot{m},\omega))$.

Fig. 5 shows a schematic diagram in Dymola showing the experiment simulated for the turbine module. The turbine module model implemented used isentropic approximations using all components from `Modelica.Fluid` and `Modelica.Media`. The gas used is dry air as ideal gas. Fig. 5 shows the inertia component representing the shaft inertia, over which a torque ramp is applied trying to simulate a disturbance.

Fig. 6 depicts the evolution of angular velocity of the shaft `omega`, and the mass flow rate `mdot_compressor`. Both are the simulated variables, under the unique boundary condition that is the driving torque `Tu_ext`. This torque emulates a brake acting on the shaft. As in the compressor case, the initial value for the angular velocity of the shaft is computed by Dymola. The torque negative ramp decrements the angular velocity under constant thermofluid boundary conditions.

3 Computational Causality And Conditions for Numerical Convergence

Dymola tool will apply several symbolic manipulation algorithms to formulate the DAE of the turbine-set, and to define the set of algebraic equations to successfully find initial conditions. This section presents a simplified set of equations for the initialization of the DAE, formerly presented and ordered by hand trying to find the minimum number of algebraic loops. That is, computational causality has been calculated manually and the variables between brackets ([ ]) represent the computed variable from that equation.

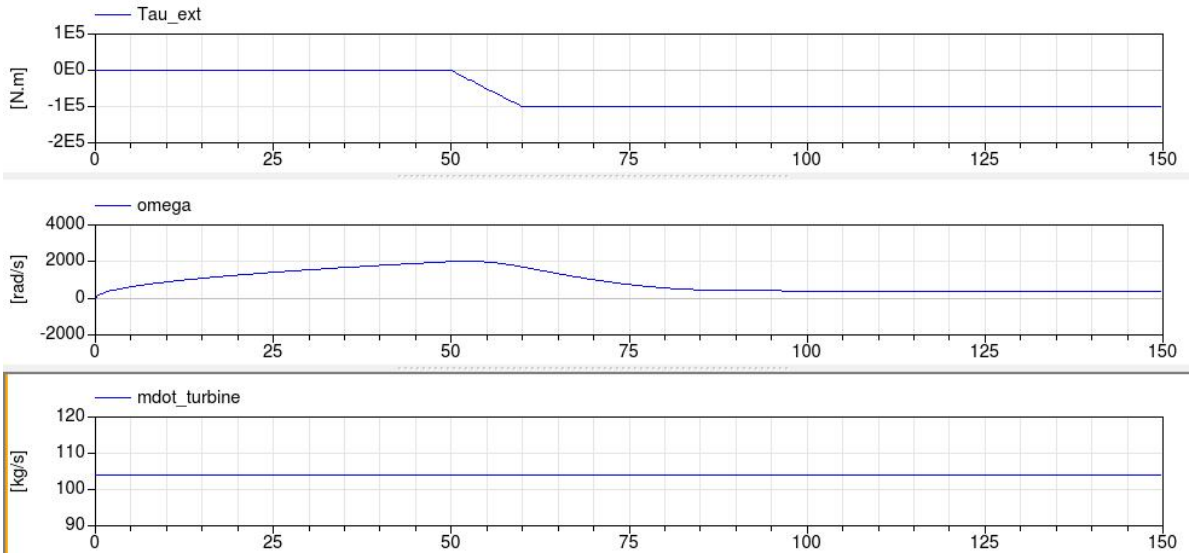


Figure 6. Simulation of the turbine under constant thermofluid boundary conditions and a ramp in the driving torque acting over the inertia.

The known variables are $\{\omega, p_{t,o}, p_{c,i}\}$ and the unknowns are $\{\dot{m}, \tau_c, \tau_t, \Delta h_c, \Delta h_t, \dot{\omega}\}$. Where subscripts 'c', 't', 'o', 'i' stand for: compressor, turbine, outlet and inlet, respectively. Parameters are: $\{\eta_{mec,c}, \eta_{mec,t}, J\}$. Properties: $\{\Phi_c, \eta_c, \Phi_t, \eta_t\}$.

The ordered list of equations:

$$\frac{p_{t,o}}{p_{c,i}} = \Phi_c([\dot{m}], \omega)\Phi_t([\dot{m}], \omega) \tag{6}$$

$$[\Delta h_c] = \eta_c(\dot{m}, \omega) \tag{7}$$

$$\dot{m}\Delta h_c = [\tau_c]\,\omega\eta_{mec,c} \tag{8}$$

$$[\Delta h_t] = \eta_t(\dot{m}, \omega) \tag{9}$$

$$\dot{m}\Delta h_t = [\tau_t]\,\omega\eta_{mec,t} \tag{10}$$

$$\tau_c + \tau_t = J\,[\dot{\omega}] \tag{11}$$

Only in the first equation from the list, (6), an algebraic loop appears involving the compressor and turbine maps. Usually is a non-linear equation that must be solved by numerical methods, although it is important to note that the maps should be tested for numerical convergence of this equation.

Actually the choice of known variables is similar to that of the dynamic initial value problem (IVP) formulated, so the testing of convergence of (6) should be mandatory.

The model presented in section 2 based in `Modelica.Fluid` and `Modelica.Media` is more complex and the causality analysis performed by Dymola could vary, but the conclusion about the necessary good numerical behaviour of compression and expansion maps applies too, beside new others.

4 First Principles Compression and Expansion Maps

4.1 Compression Case

Compression maps are usually obtained from steady state measurements or data from manufacturer. When those maps are not available, first principle solutions may be implemented. In (Greitzer, 1976; Gravdahl and Egeland, 1997, 1999; Gravdahl et al., 2000, 2004) a description of the physical principles used in the deduction of a generic compression map are presented and more references for deeper details can be found. Based on these, may be concluded that Δh_s in (4) could be approximated under some assumptions by (12).

$$\Delta h_s = \sigma r_2^2\omega^2 - \frac{r_1^2}{2}(\omega - \alpha\dot{m})^2 - k_f\dot{m}^2 \tag{12}$$

Under isentropic conditions, the main contribution of this formulation is that Δh_s is formulated as an algebraic relation with a second-degree polynomial in $\dot{m}$, that let us easily rearrange (12) to obtain an explicit relation for $\dot{m}$, obtaining (13). This choice avoids non linear algebraic loops in (6).

$$\dot{m} = \frac{\alpha\omega r_1^2 \pm \sqrt{D(b,\omega)}}{2a_2} \tag{13}$$

where:

$$D(b,\omega) = \alpha^2\omega^2 r_1^4 - 4a_2\left(a_{01}\omega^2 + b\right) \tag{14}$$

$$b(\Phi_c, c_{pi}, T_i) = \left(\Phi_c^{\frac{\kappa-1}{\kappa}} - 1\right) c_{pi}T_i \tag{15}$$

In (12), (13) and (14) a new set of six parameters appears: $\{\sigma,\ \alpha,\ r_1,\ r_2,\ k_f,\ a_{01},\ a_2\}$ where only the first

five are independent, being $\{a_{01}, a_2\}$ dependent of them. Please, note that b depends on the boundary conditions for the case of a compressor model. For more details, please read section 7.

The conditions for model validity are defined in equations (16) and (17), and the model will be singular if unfulfilled any of both. This is convenient to consider in the initialization code.

$$\omega \geq \omega_{min}(b) = 2\sqrt{\frac{a_2 b}{\alpha^2 r_1^4 - 4a_2 a_{01}}} \tag{16}$$

$$\dot{m} \geq \dot{m}_{min}(b) = \frac{\alpha \omega_{min} r_1^2}{2a_2} \tag{17}$$

4.2 Expansion Case

For the turbines, similar equations should be derived. A typical extended case is the Stodola turbine model, with results presented in section 2.3, in a simplified way to avoid the general approach described in section 3, expressed by (18) and (19).

$$\dot{m} = K\sqrt{\frac{p_i}{T_i}(1-\Phi_c^2)} \tag{18}$$

$$\eta_t = \eta_t(\Phi_c, \omega, T_i) \tag{19}$$

Where K is a parameter. These explicit expressions of $\dot{m}$ and η_t avoid the algebraic loops previously referred.

5 Simulation Results

The composition diagram of the compressor-turbine system Modelica model is shown in Fig. 7. For the compressor model, the above explained first principles approximation for the compressor map from (12) to (17) has been implemented. For the turbine module, the Stodola model has been applied based on (18) and (19).

The experiment which results are shown in Fig. 8 consists of applying a sequence of ramps for input heat power to the combustor, equivalent to gas mass flow rate ramps. A step for external torque is applied from the beginning, emulating a *cold* start-up to reach initial conditions inside the validity region of the model. These ramps are represented in `u_ThermalPowerPercent` and `u_TorquePercent` as variables representing the percentage of the total thermal power injected and the torque applied to the axis. The variables returned from the simulation are: `f` the frequency of the axis, `mdot_compressor` the mass flow rate of air through the compressor and `T_Combustor` the temperature inside combustor. It can be observed a (non-linear) first order behavior for the three simulated variables when assuming the input ramps as quasi-steps, due to the short up/down times. Different time constants and steady state gains must be found for each operation point and step applied. As shown, `T_Combustor` rises at the time of torque negative edge due to the decreasing change in mass

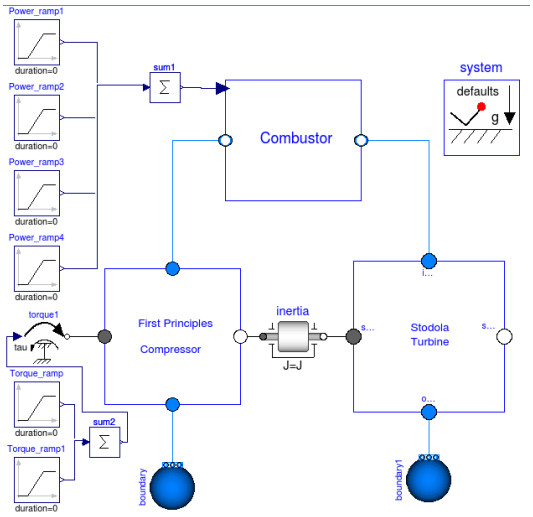


Figure 7. Diagram of the compressor-turbine system in Dymola tool. First principles compressor map and Stodola turbine models have been used. For both inputs (heat flow into combustor and torque for *cold* start-up) a combination of ramps are used.

flow rate, that increases residence time of fluid in the combustion chamber. This figure shows the exhaust flue gases temperature `T_Out_Turbine`, which is different from the temperature in the combustor.

From a control point of view, due to the restrictions of the model validity region imposed by the presented equations, only input thermal power may be used as manipulated variable and the torque may be used as disturbance.

6 Conclusions

A gas turbine dynamic model composed of different components has been presented. The main objective of the model is to be used in real time simulations and control applications. Only the main dynamics of a gas turbine need to be predicted in a wide operation range and with the minimum complexity to obtain a dynamic inverted model to be used inside the controllers. An analysis is presented on computational causality in the initialization and the IVP with concluding requirements for numerical convergence. A proposal for compression and expansion maps, derived from first principles are presented, to be used when unavailability of data from manufacturer and avoiding the formation of algebraic loops. Some simulation experiments are performed and the results shown.

Acknowledgment

The authors gratefully acknowledge the funding support from CIEMAT Research Centre, EU 7^{th} Framework Programme (Theme Energy 2012.2.5.2) under grant agreement 308912 - HYSOL project - Innovative Configuration of a Fully Renewable Hybrid CSP Plant, the National R+D+i Plan Project DPI2014-56364-C2-2-R of the Spanish Ministry of Economy and Competitiveness and ERDF funds.

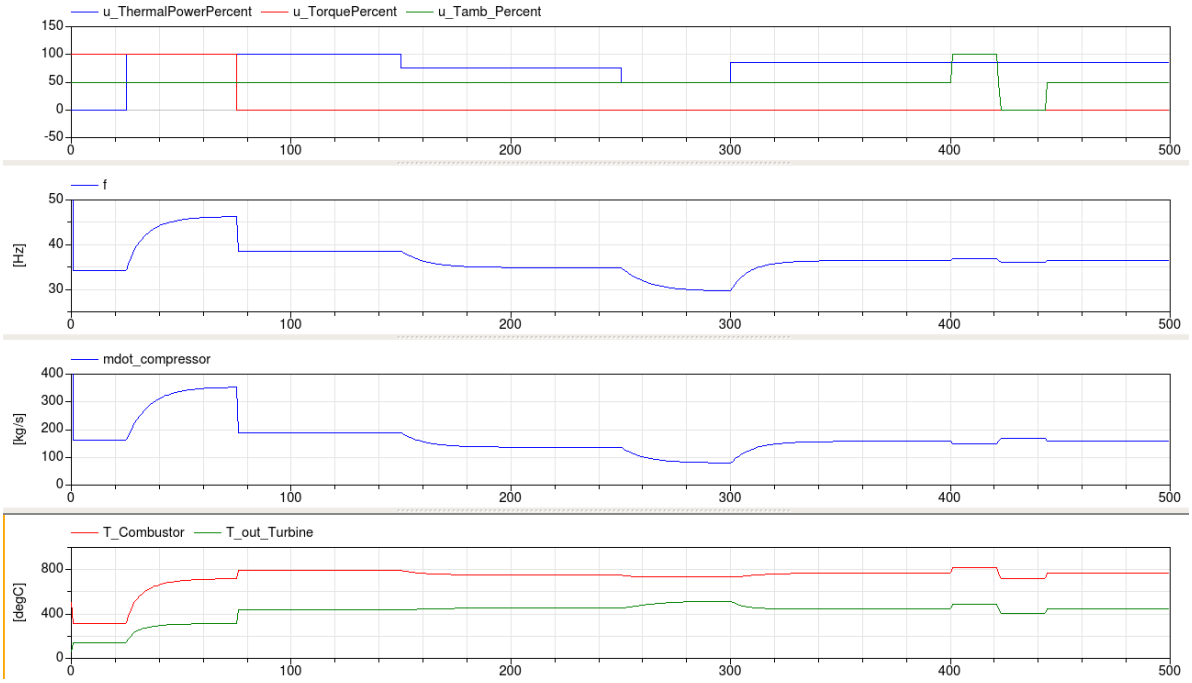


Figure 8. Input variable and simulation results for the model in Fig. 7.

References

3DS. *Dymola 2016 User Manual*, 2016.

W. W. Bathie. *Fundamentals of gas turbines*. Wiley, second edition, 1996. ISBN 9780471311225.

F. Casella and A. Leva. Modelling of thermo-hydraulic power generation processes using Modelica. *Mathematical and Computer Modelling of Dynamical Systems*, 12(1):19–33, 2006. doi:10.1080/13873950500071082.

F. Casella, M. Otter, K. Proelss, C. Richter, and H. Tummescheit. The modelica fluid and media library for modeling of incompressible and compressible thermo-fluid pipe networks. In *Proceedings of the Modelica Conference*, pages 631–640, 2006.

F. E. Cellier. *Continuous System Modeling*. Springer-Verlag, 1991.

H. Elmqvist, H. Tummescheit, and M. Otter. Object-oriented modeling of thermo-fluid systems. In *3rd International Modelica Conference*, pages 269–286, 2003.

P. Fritzson. *Principles of object-oriented modeling and simulation with Modelica 2.1*. Wiley-IEEE Press, 2004.

J. T. Gravdahl and O. Egeland. Moore-greitzer axial compressor model with spool dynamics. In *Proceedings of the IEEE Conference on Decision and Control*, volume 5, pages 4714–4719. IEEE, 1997.

J. T. Gravdahl and O. Egeland. Centrifugal compressor surge and speed control. *IEEE Transactions on Control Systems Technology*, 7(5):567–579, 1999. ISSN 10636536. doi:10.1109/87.784420.

J. T. Gravdahl, F. Willems, B. De Jager, and O. Egeland. Modeling for surge control of centrifugal compressors: Comparison with experiment. In *Proceedings of the IEEE Conference on Decision and Control*, volume 2, pages 1341–1346, 2000.

J. T. Gravdahl, F. Willems, B. De Jager, and O. Egeland. Modeling of surge in free-spool centrifugal compressors: Experimental validation. *Journal of Propulsion and Power*, 20(5): 849–857, 2004. ISSN 07484658.

E. M. Greitzer. Surge and rotating stall in axial flow compressors. part i: Theoretical compression system model. *Journal of Engineering for Gas Turbines and Power*, 98(2):190–198, 1976.

S. C. Gülen and K. Kim. Gas turbine combined cycle dynamic simulation: A physics based simple approach. *Journal of Engineering for Gas Turbines and Power*, 136(1), 2014.

R. Kehlhofer, B. Rukes, F. Hannemann, and F. Stirnimann. *Combined-cycle gas & steam turbine power plants*. Pennwell Books, 2009.

S. V. Patankar. *Numerical Heat Transfer and Fluid Flow. Series in Computational and Physical Processes in Mechanics and Thermal Sciences*. Taylor & Francis, Mortimer House, 37-41 Mortimer Street, London, W1T 3JH, 1980.

K. J. Åström, H. Elmqvist, and S. E. Mattsson. Evolution of Continuous-Time Modeling and Simulation. In R Zobel and D Moeller, editors, *Proceedings of the 12th European Simulation Multiconference, ESM'98*, pages 9–18, Manchester, UK, June 1998. Society for Computer Simulation International.

H. Tummescheit. Design and Implementation of Object-Oriented Model Libraries using Modelica. 2002.

H. K. Versteeg and W. Malalasekera. *An Introduction to Computational Fluid Dynamics*. Addison Wesley Longman Limited, Pearson Education. Edinburgh Gate. Harlow. CM20 2JE. United Kingdom., 1995.

7 Appendix

From (Gravdahl and Egeland, 1999) and its references, the remaining equations and parameters for the model are exposed below. All the parameters are in table 1. Slip factor σ in (12) is a parameter that depends on the construction but is assumed to be minor but close to 1. The expression used is (20). Parameter α is defined by (21).

$$\sigma = 1 - 2/N_b \tag{20}$$

$$\alpha = \frac{Acot(\beta_{1b})}{A_i r_1 \rho_i} \tag{21}$$

The total friction factor k_f in (22) is the overall friction factor from all friction losses.

$$k_f = \frac{4fl}{2D\rho_i^2 A_i^2 sin^2(\beta_{1b})} \tag{22}$$

For friction factor expression any of the bibliography can be used, although the Blasius' formula is one of the most frequently used (23). In this case, f has been considered a parameter for the sake of simplicity.

$$f = 0.3164 Re^{-0.25} \tag{23}$$

The dependent parameters $\{a_{01}, a_2\}$ are defined by (24) and (25).

$$a_{01} = \frac{r_1^2}{2} - \sigma r_2^2 \tag{24}$$

$$a_2 = \frac{\alpha^2 r_1^2}{2} + k_f \tag{25}$$

Table 1. Models Parameters

Parameter	*Magnitude*
N_b	Number of compressor blades
f	Friction factor
l	Mean channel length in the compressor
D	Mean hydraulic diameter for the compressor
A	Section of the circle with diameter D
A_i	Inlet section of the compressor
ρ_i	Inlet air density
β_{1b}	Backswept inlet vane angle
r_2	Impeller outer radius
r_1	Impeller inner radius

Object-Oriented Modelling and Simulation of a Molten-Salt Once-Through Steam Generator for Solar Applications using Open-Source Tools

Francesco Casella Stefano Trabucchi

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy,
{francesco.casella,stefano.trabucchi}@polimi.it

Abstract

Concentrated solar power plants (CSP) coupled with thermal storage have the potential of being competitive with conventional fossil fuel and hydro power plants, in terms of dispatchability and provision of ancillary services. To achieve this potential, the plant design has to be focused on flexible operation, which is the main goal of the PreFlexMS Horizon 2020 European research project. This can be achieved by the integration of a Molten Salt Once Through Steam Generator within the power unit, an innovative technology with greater flexibility potential if compared to steam drum boilers, currently the state of the art in CSP. Given the focus on flexible operation, dynamic modelling and simulation from the early design stages is of paramount importance, to assess the plant dynamic behaviour and controllability, and to predict the achievable closed-loop dynamic performance, potentially saving money and time during the detailed design, construction and commissioning phases. The present paper aims to demonstrate that it is possible to achieve this goal by means of Modelica-based open-source libraries and tools.

Keywords: simulation, Modelica, solar power plants, once-through boilers, open-source tools.

1 Introduction

During the 2015 United Nations Climate Change Conference 195 countries agreed to reduce their green-house gases production "as soon as possible", in order to keep the global warming below 2 °C . Research on Concentrated Solar Power (CSP) plants is therefore of paramount importance, since such plants could be able to fill the gap between renewable and fossil energy sources, in terms of predictability, flexibility and power production planning.

In this context, the PreFlexMS Horizon 2020 research project aims to design a 100 MW_{el} CSP power unit with flexibility features comparable to state-of-the-art gas-fired combined cycle plants, and to demonstrate it on a scaleddown pilot plant. The two pillars of this project are: the optimized operation of the plant based on customized weather prediction and on the demand on the electrical market, on one side, and a power unit based on an innovative once-through molten-salt steam generation unit on the other side. The latter is expected to provide much more flexible operation than the current state of the art in CSP, i.e., steam-drum type steam generators.

Dynamic modelling and simulation is a key aspect of the project. The plant dynamic behaviour and controllability need to be assessed from the early design stages, and it is essential to demonstrate that the achievable closedloop performance of the plant is consistent with the project goals, which ask for: a) an overall hot start-up time of 30 minutes and b) ramping rates comparable to state-of-the-art combined-cycle plants.

The goal of this paper is to describe the modelling approach that was taken in this project, and in particular to demonstrate that object-oriented open standards and open source tools are perfectly suited to carry out this task.

The paper is structured as follows. Section 2 introduces the modelling methodology, language, model libraries and simulation tools that were used in the project. The plant subject of the study is described in Section 3; as to its modelling, Section 4 discusses new device models that were created for this project, while Section 5 discusses the overall plant model. The simulation objectives and methodology is presented in Section 6, while simulation results are presented in Section 7. Finally, Section 8 concludes the paper.

2 Modelling methodology, libraries and tools

The equation-based, object-oriented modelling approach (EOOM) and the Modelica language have been employed for this work. A general discussion of this topic goes beyond the scope of this paper, the interested reader is referred to (Mattsson et al., 1998; Tiller, 2001; Fritzson, 2014). We only stress two key points here: on one hand, that dynamic models are described by means of declarative descriptions, based on differential-algebraic equations, and of modular composition through a-causal physical connectors; on the other hand, that the Modelica language used to code the model is a non-proprietary, tool-independent standard.

A key feature of the EOOM is that it promotes reusability, while at the same time allowing for extensions and customizations with a minimal effort. This was also

the case for the present study, which relied on existing open-source model libraries. Besides the Modelica Standard Library, which contains utility models as well as the strategically important IAPWS IF97 water/steam model, the ThermoPower and the IndustrialControlSystem Modelica libraries were used. The first one is a library for the modelling of thermal power plants, whose design principles are discussed in (Casella and Leva, 2006) and that has been used, adapted and validated extensively to model various types of power generation and energy conversion systems such as steam generators (Casella and Leva, 2005), combined-cycle power plants (Casella and Pretolani, 2006; Bartolini et al., 2012), nuclear power plants (Cammi et al., 2011), cryogenic systems (Zanino et al., 2013), and Organic Rankine Cycle systems (Casella et al., 2013). The second one is the IndustrialControlSystem library (Bonvini and Leva, 2012), containing detailed models of industrial PID controllers endowed with features such as anti-windup, output tracking, etc., that were needed for the modelling of the control system.

Device models not found in ThermoPower had to be adapted or written from scratch, see Section 4.

In principle, any Modelica-compliant simulation tool can be used to analyze and simulate it. In this paper, we discuss the results obtained by using the open-source OpenModelica tool, to demonstrate that the dynamic modelling and simulation task can be carried out using a completely open-source tool chain.

3 Plant Description

The subject of this study is an innovative 100 MW_{el} Concentrated Solar Power (CSP) plant with thermal storage. The solar field directs the solar radiation onto the receiver tower, where a mixture of Sodium and Potassium nitrate (commonly called hereafter molten salt (MS)) is heated up to 565 °C and stored in a tank. The molten salt (MS) acts both as Heat Transfer Fluid (HTF) and as Thermal Energy Storage (TES). In the context of this study, whose scope is limited to the power generation block, the hot MS is considered as coming from a source point at constant temperature.

The model represents the power unit, composed of the steam generator system (SGS), the steam turbine, the feedwater heaters train with a deaerator, and the plant control system. The SGS is the core of the dynamic analysis, while the turbine and the feedwater heaters provide the correct boundary conditions for the boiler.

3.1 Innovative aspects

The main innovative aspect of the present plant is the introduction of a Molten Salt Once-Through Steam Generator (OTSG) instead of a drum-type boiler, which represents nowadays the state of the art concerning CSP with thermal storage. Such a system can provide a higher plant flexibility, since the drum and its related large inertia and thermo-mechanical stresses are absent.

The OTSG is a continuous flow path of Shell&Tube heat exchangers, whose mechanical design is optimized in order to maximize the life-cycle of the whole boiler. The MS flow is cooled down to the design value of 290 °C while on the other side superheated steam is produced.

Another non-conventional issue is represented by the hot fluid thermal inertia. Compared to once-through boilers used in advanced combined-cycle (CC) plants, where the thermal inertia of the hot fluid is negligible, the heat capacity of the MS in this plant is the dominant thermal storage in the process, so that the thermal coupling between the MS and the water/steam side and in general the plant dynamic response is quite different from that of existing CC plants.

3.2 OTSG topology

As the plant is meant to operate in sliding pressure mode, the joint effect of evaporation and bleed steam pressure reduction would significantly decrease the MS OTSG outlet temperature and the feedwater (FW) temperature at the same time, during part-load operation. This phenomenon is not acceptable, because it could lead to MS freezing, which is a catastrophic fault. Thus, the process is designed so that the MS outlet temperature is kept almost constant as the plant load is reduced.

The key aspects of the heat exchangers topology are the following (see Figure 1):

- the greater amount of feedwater mass flow enters the OTSG at the economizer (ECO) cold end, flows through the evaporator (divided into EVA1 and EVA2) and enters a phase separator. If the steam quality at the EVA2 hot end is equal to one, then the separator does not affect the steam thermodynamic condition; otherwise, the saturated steam enters the superheating section (SH) and the saturated liquid is collected at the separator bottom and recirculated at the OTSG inlet;
- a small amount of feedwater mass flow enters a MS heater, and is then recirculated to the last high pressure feedwater heater cold end;
- from the MS line point of view, the steam SH and reheating (RH) sections run in parallel: the hot MS mass flow entering the OTSG is split up in two branches by means of a valve, and then mixed at the evaporation section hot end;
- a by-pass valve on the last high pressure feedwater heater offers another degree of freedom, useful to obtain the desired feedwater temperature at OTSG inlet.

Contrary to the bottoming cycle of a modern combined cycle, feedwater heaters are needed in order to provide a sufficiently high water temperature at the OTSG inlet and to prevent the MS from freezing.

The result is a non-conventional steam Rankine cycle, with a high number of recirculation flows interacting in a non-trivial way, thus making the dynamic simulation a key factor during system and control design.

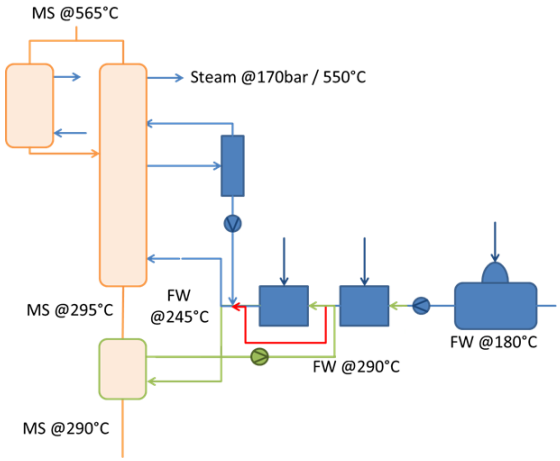


Figure 1. OTSG and high pressure feedwater train topology, patent pending EP 15290109.6-1610

4 Device modelling

As a first step, all the physical devices of the system are modelled, by means of both re-used and custom models.

4.1 Re-used components

The dynamic properties of a thermal power system are mainly given by the heat exchangers design, in terms of heat exchange surface, volumes and metal masses. These parameters are set within the *Flow1DFV* component of the ThermoPower library, which represents by means of 1D finite volumes method the hot and cold paths of each heat exchanger. The two equivalent tubes are connected through the model of a thin metal wall, where the heat exchanger mass is concentrated. Finally, the *heatExchangeTopology* component lets the user set the flow path displacement, such as co-current, counter-current and Shell&Tube (Boni, 2013).

Other simple components have been taken directly from the ThermoPower library, such as liquid and steam valves, pressure drops, and pumps, and they are connected between each others through standard Modelica stream connectors (Franke et al., 2009). The water/steam thermodynamic properties are given by the IF97 tables of the Modelica standard library.

4.2 Custom component models

The complexity and non-standard features of the power unit required the development of custom models of specific components, in order to represent in the best way possible their physical behaviour.

4.2.1 Separator

At the end of the evaporation section, a two-phase cyclone separator guarantees that only saturated or superheated steam enters the superheating section, so that no liquid drops can reach the turbine inlet. Since from the geometrical point of view this device is a high and narrow cylinder with a small cross sectional area and the inlet flange positioned in the upper part, it can be assumed that the thermal contact between the liquid possibly present in the lower part and the wet/dry steam is negligible.

The four main equations describing the dynamics of the separator are thus the mass and energy balances over the liquid and steam volumes, assuming as states the liquid level l, the liquid enthalpy h_l, the pressure p and the vapour enthalpy h_v.

The outlet steam condition can be either saturated or superheated, depending on the steam quality at the evaporators outlet. The following equation guarantees that no wet steam enters the SH:

$$h_{v,out} = max(h_v, h_{v,sat}(p)) \tag{1}$$

4.2.2 Feedwater heater

A feedwater heater is a particular heat exchanger where the steam extracted from the turbine condenses on a tubes bundle. The feedwater flow inside the bundle sets the amount of steam that actually condenses: the lower the feedwater mass flow, the lower the condensing mass flow.

The mechanical design of such an heat exchanger usually assumes three different heat exchange zones:

1. a desuperheating section, where the bleed steam from the turbine is brought to saturation condition. Since the bled steam is not always superheated, this zone could be missing, especially in low pressure heaters train;

2. a condensing section, where the bleed steam and the drained subcooled liquid coming from the higher pressure heaters mix up and condense;

3. a subcooling section, where the condensate is subcooled.

Such a structure has been modelled as three heat exchangers in series, once more making use of the component *Flow1DFV* of the ThermoPower library, as shown in Figure 2. A valve downstream the steam subcooling section controls the condensate level inside the condensing section by means of a PI controller, discharging the drained flow into the next, lower pressure heater.

The key base component of the feedwater model is the steam condenser, called *NusseltCondenser*. It is modelled as a cavity where the steam fraction coming from the drained and the bleed flows condenses. As for the separator, the four main equations describing the dynamics of the separator are the liquid and steam mass and energy balances, assuming as states the liquid level l, the liquid enthalpy h_l, the pressure p and the vapour enthalpy h_v. Thanks to the *min* and *max* operators and assuming that the liquid entering the cavity goes directly into the liquid pool at the shell bottom, the energy balances can cover also the cases where the entering steam is wet, usually at part load.

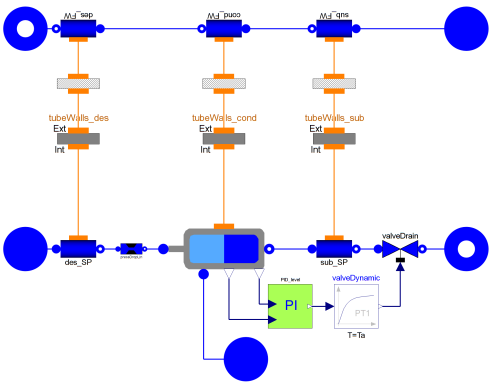


Figure 2. Object diagram of a feedwater heater

The tube bundle on which steam condensation takes place is divided into N_w finite volumes, as the *Distribuited Heat Transfer (DHT)* connectors in Figure 2 shows. This means that the total condensate mass flow and heat flux is the sum of the single contributions of each volume. The equations for the $i-th$ volume read:

$$Q_i = max\left[0, \frac{\gamma S}{N_w}\left(T_{sat} - T_{c,i}\right) + w_{c,i}\left(h_v - h_{v,sat}\right)\right] \quad (2)$$

$$w_{c,i} = \frac{Q_i}{h_v - h_{c,i}} \quad (3)$$

$$\text{condState}_i = \text{setState\_}pT\left(p, T_{c,i}\right) \quad (4)$$

$$h_{c,i} = min[\text{specificEnthalpy}(\text{condState}_i), h_{l,sat}] \quad (5)$$

$$T_{c,i} = \frac{T_{w,i} - T_{sat}}{2} \quad (6)$$

where the subscript "c" refers to the condensate condition and "w" to the metal tubes. Thus:

$$Q_{flux} = \sum_{i=1}^{N_w} w_{c,i} h_{c,i} \quad (7)$$

$$Q_{cond} = \sum_{i=1}^{N_w} Q_i \quad (8)$$

$$w_{cond} = \sum_{i=1}^{N_w} w_i \quad (9)$$

where Q_{cond} refers to the heat power entering the metal wall, w_{cond} the actual condensate flow and Q_{flux} the global enthalpic flux entering the liquid pool through by means of the condensate.

During some transients the liquid enthalpy could become greater than the saturated liquid enthalpy, so that a little amount of liquid evaporates. Hence, the following term has been added within the mass balances:

$$w_{ev} = \frac{x_l \rho_l V_l}{\tau_{ev}} \quad (10)$$

where x_l is the steam fraction in the liquid volume and τ_{ev} is an equivalent rising time, a tuning parameter.

4.2.3 Deaerator

The deaerator is a component of paramount importance within a steam cycle. Its main function is to extract from the feedwater flow the incondensable gases, which would chemically react at high temperature compromising the lifetime of high-temperature components such as the SH and the high pressure turbine.

It is worth noticing that the presence of incondensable gases and the steam blown away through the discharge valve can be omitted entirely from the model, due to their very modest quantities, while the hydraulic and thermal effects the deaerator has on the rest of the cycle must be taken into account, due to its large size.

The most efficient working condition for such a component is the thermodynamic equilibrium between steam and water, since it promotes the gases dissolution from feedwater mass flow. Under the assumption that the design and the local control system of this device guarantee the optimal condition, it can be modelled as a parallelepiped-shaped vessel at thermodynamic equilibrium, where all the entering flows mix up.

Assuming as states the pressure p and the mixture enthalpy h, the main model equations are the mass and energy balances on the device. The liquid level can be derived from the steam quality, as follows:

$$l = H\left(1 - \frac{\rho}{\rho_{v,sat}} x\right) = H\left(1 - \frac{\rho}{\rho_{v,sat}} \frac{h - h_{l,sat}}{h_{v,sat} - h_{l,sat}}\right) \quad (11)$$

4.2.4 MS fluid model

The MS has been represented as an extension of the *Modelica.Media.Interfaces.PartialPureSubstance* class. References value for enthalpy, density and temperature have been set according to the specific MS thermo physical properties. Density and specific heat capacity are considered as linear functions of the temperature.

5 Plant modelling

The complete plant model is composed by the four macro components described in Sec. 3. Thanks to the *encapsulation* feature of the Modelica language, each component can be easily replaced with more or less detailed ones, without modifying the rest of the system. Since there are several possible plant operational modes that have been tested, this aspect hugely reduced the time spent for test cases setup.

5.1 OTSG model

The OTSG model follows the process flow diagram of Figure 1. Six heat exchanger models are connected via stream connectors, within a base class model that contains the common OTSG fluid and control signal interfaces.

Two important aspects concerning the pressure drops representation have to be highlighted:

- the MS pressure drops are neglected. As the MS is always in liquid state, its density and specific heat capacity are hardly influenced by the fluid pressure;

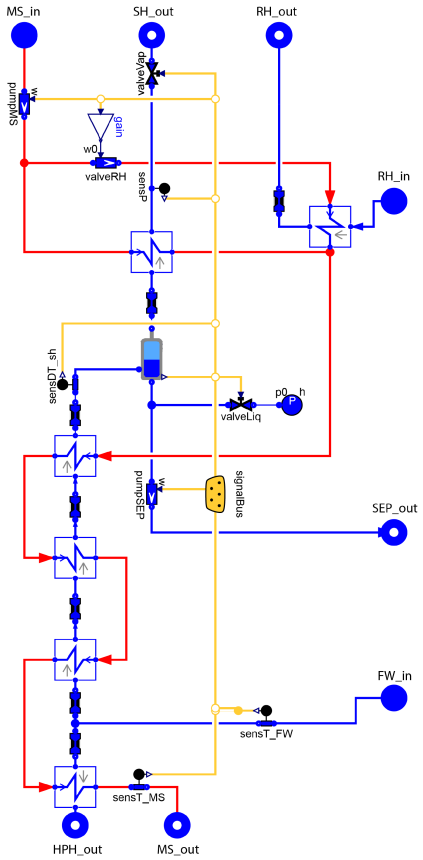


Figure 3. Object diagram of the OTSG

- the steam/water pressure drops are taken into account through the ThermoPower component *Water.PressDrop* at the inlet or the outlet of each heat exchanger. This approach have two main advantages: first, the coupling between mass and energy balances is reduced, so that the resulting algebraic loops are easier to be solved; secondly, each heat exchanger is independent from the equations system point of view, so that the resulting symbolic manipulation produces a more efficient simulation code.

The blue lines in the figure refer to the water/steam flows, the red lines to the MS flows and the yellow lines to the control signals. The water/steam pressure drop components are the blue rectangular boxes in between the heat exchangers, while the gain block splits the MS mass flow as required between SH and RH branches. Heat exchangers and flow lines parameters have been taken from design Process Flow Diagrams (PFD) and heat exchangers datasheets.

5.2 Turbine model

The turbine model was built by connecting several ThermoPower components *Water.SteamTurbineStodola*, each representing a turbine section with many stages, and accounting for feedwater heater steam extractions and for leakage and sealing flows. The model parameters (Stodola's law flow coefficient and isentropic efficiency) have been extrapolated from the PFDs. Finally, all the turbine sections are connected through mechanical interfaces representing the common turbine shaft.

5.3 Feedwater preheaters train

The power plant layout features a high pressure and a low pressure feedwater heater train, with the deaerator in between. The feedwater models have been connected to reproduce the topology reported in Sec. 3.

5.4 Plant control system

Once the single devices have been tuned and tested, the overall open-loop system dynamics has been thoroughly analysed, in order to develop a control architecture and controller tunings that were able to meet the control requirements.

More specifically, the model was numerically linearized at different operating load levels. The linearized models allowed to compute step responses, input-output transfer functions and frequency responses (Bode diagrams), as well as the Relative Gain Array (RGA) matrix, which allowed to investigate the couplings between the different control loops. This analysis led to the definition of a decentralized control strategy based on PID controllers and static feed-forward set-point compensation. Figure 4 shows the block diagram of one of these controllers, which are also defined by Modelica code.

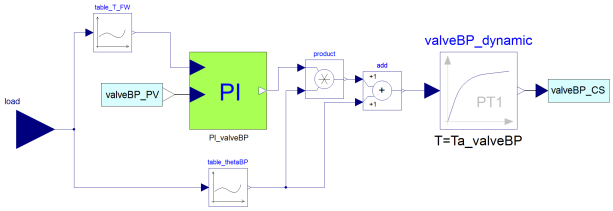


Figure 4. Object diagram of the high-pressure feedwater heaters by-pass control loop

5.5 Complete plant model

All the macro components can be aggregated in a single global system model, representing the whole power unit of the CSP plant. In Figure 5 the complexity of the system is evident, thus only the high-pressure part of the plant has been considered. This simplification does not affect the analysis results: the big size of the deaerator, in fact, dynamically decouples the high pressure from the low pressure part of the plant.

6 Simulation objectives - methodology

The simulation campaign aims to demonstrate that such a complex system can be simulated with open-source tools (e.g. OpenModelica). In particular, three different test cases have been analysed:

- *A - nominal load*: simulation of the nominal plant regime, to check the static correctness of the model;

TABLE 1. Machine Features

	Laptop	Workstation
CPU	Intel(R) Core(TM) i7-4810MQ	Intel(R) Xeon(R) E5-2650 v3
Clock	2.8 GHz	2.3 GHz
RAM	8 GB	72 GB

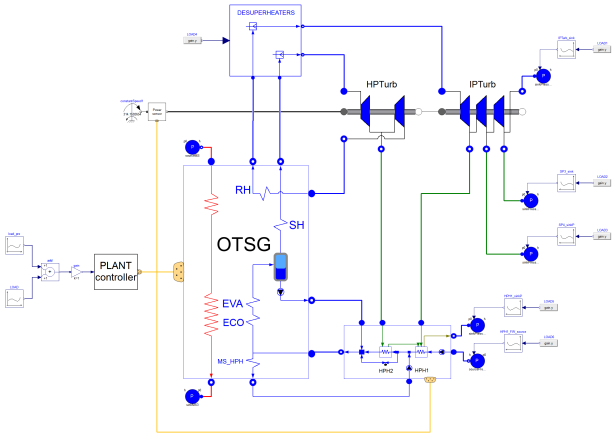


Figure 5. Object diagram of the high-pressure part of the power unit, used for daily start-up simulation

- *B - load ramps*: simulation of a series of load ramps to check the precision and robustness of the developed control system. The model is initialized at 100% load, brought to 60% with a slow negative ramp and then back to 100% by ramps having a 10% amplitude and a 10%/min slope;
- *C - plant hot start-up*: simulation of the daily power plant hot start-up. For numerical reasons, the model is first initialized at 40% load and brought to the initial load level of 20% with a slow negative ramp. Then, during the actual start-up simulation the load is brought back to 100% in 9 minutes.

The models have been compiled and simulated on a Linux workstation and on a Windows laptop, whose features are reported in Table 1. The sizes of the three models are summarized in Table 2.

7 Simulation performance

The thorough description of the simulated transients goes beyond the scope of this paper. For the sake of the example, we report the total mechanical power output and the temperatures of the feedwater and molten salt at the OTSG cold end, which should never go below 245 °C during the hot start-up transient. The power set-point is closely followed by the actual value, with a small glitch when the operating mode of the evaporator is switched from wet operation with recirculation to dry, once-through mode. The MS temperature is always above the safety lower limit and the feedwater temperature closely follows its set-point (Figure 6).

TABLE 2. Test case models information

Case	no. States	no. Variables
A	254	6945
B	254	6982
C	301	7554

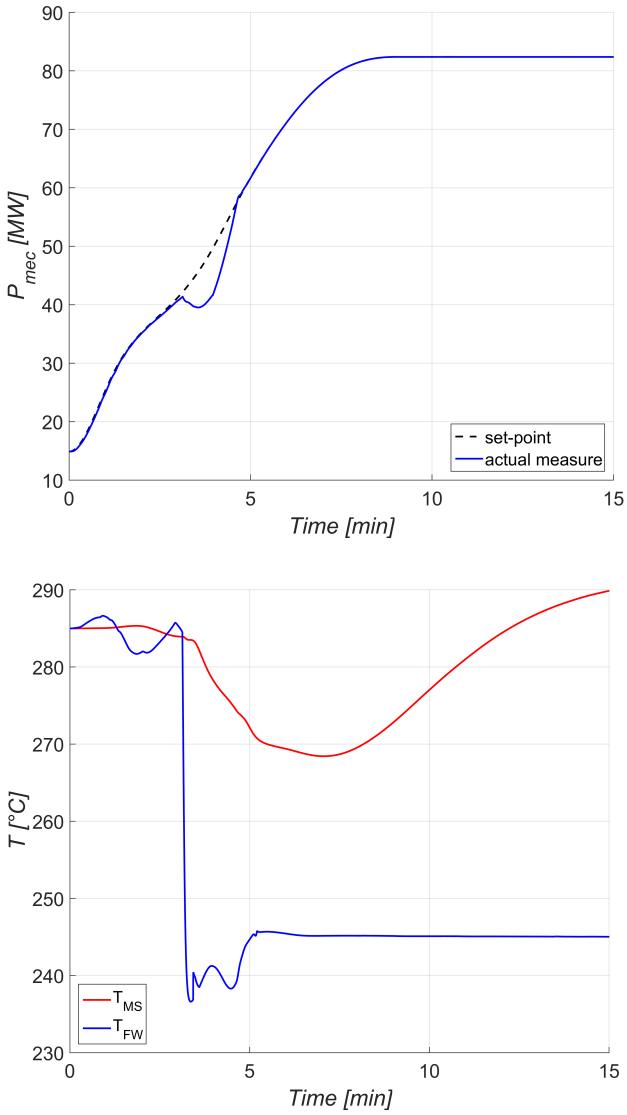


Figure 6. Mechanical power, MS and feedwater temperatures during hot start-up

TABLE 3. Simulation performance

Case	CPU compilation time [s]		no. integration steps		CPU integration time [s]	
	Linux	Windows	Linux	Windows	Linux	Windows
A	5.7	32	2898	3062	161	170
B	5.5	19	11030	10768	670	636
C	6.7	22	6404	6635	410	390

The test campaign confirms the capability of OpenModelica to compile and simulate correctly complex models with a high number of states and variables. Table 3 reports the simulation performance for all the three test cases considered. The numerical algorithm employed is DASSL, without equidistant grid result interpolation, and with a relative tolerance of 10^{-6}.

8 Conclusions

The results presented in this paper demonstrate that it is possible to tackle the detailed, system-level dynamic modelling and simulation of innovative power generation system for use in solar thermal power plants, using exclusively open modelling languages and simulation tools.

Modelica was chosen as the modelling language, since it is a tool-independent language that can also be used with open-source simulation tools, and for which many open-source modelling libraries are currently available.

The plant model was developed by extensively re-using basic models from the open-source ThermoPower library, with some adaptations and extensions. This model was first used to carry out extensive open-loop dynamic analysis, which allowed to design appropriate control strategies. The control system model was then also implemented in Modelica, using the IndustrialControlSystems open-source library, allowing to validate the closed-loop performance in non-trivial scenarios such as the hot start-up of the plant.

Both open-loop and closed-loop transients have been simulated by means of the open-source OpenModelica tool. The simulation performance is satisfactory: the simulation code is compiled from the Modelica source code in a few seconds, and the most demanding closed-loop transient is simulated in about ten minutes. Most of the simulation time (about 75%) is actually spent computing the water-steam properties, since the accurate but computationally demanding IAPWS IF97 model is used for that purpose. Given the on-going developments of the OpenModelica tool, it is expected that much better performance could be achieved in the near future.

Last, but not least, the choice of open-source tools also allowed to freely share the model among the different partners of the PreFlexMS Horizon 2020 project, both in source-code form and as compiled executables, without any constraints due to software licensing issues.

Acknowledgment

This work has been supported by the European Commission under Grant Agreement 654984-PreFlexMS, within the Horizon 2020 research programme.

References

Andrea Bartolini, Francesco Casella, Alberto Leva, and Luca Savoldelli. Object-oriented simulation for primary reserve scheduling in a combined cycle power plant. In *Proceedings of the 2012 IEEE Multi-conference on Systems and Control*, pages 764–769, Dubrovnik, Croatia, 3–5 Oct 2012. IEEE. ISBN 978-1-4673-4505-7.

Stefano Boni. Sviluppo di una libreria di componenti riconfigurabili per la simulazione di impianti ORC. Master's thesis, Politecnico di Milano, 2013.

Marco Bonvini and Alberto Leva. A Modelica library for Industrial Control Systems. In *Proceedings 9th International Modelica Conference*, pages 477–484, Munich, Germany, Sep. 3–5 2012. The Modelica Association. doi:10.3384/ecp12076477.

Antonio Cammi, Francesco Casella, Marco Enrico Ricotti, and Francesco Schiavo. An object-oriented approach to simulation of IRIS dynamic response. *Progress in Nuclear Energy*, 53(1):48–58, Jan. 2011. doi:10.1016/j.pnucene.2010.09.004.

Francesco Casella and Alberto Leva. Object-oriented modelling & simulation of power plants with Modelica. In *Proceedings 44th IEEE Conference on Decision and Control and European Control Conference 2005*, pages 7597–7602, Seville, Spain, Dec. 12–15 2005. IEEE, EUCA. ISBN 0-7803-9568-9.

Francesco Casella and Alberto Leva. Modelling of thermo-hydraulic power generation processes using Modelica. *Mathematical and Computer Modeling of Dynamical Systems*, 12 (1):19–33, Feb. 2006. doi:10.1080/13873950500071082.

Francesco Casella and Francesco Pretolani. Fast start-up of a combined-cycle power plant: a simulation study with Modelica. In Christian Kral, editor, *Proceedings 5th International Modelica Conference*, pages 3–10, Vienna, Austria, Sep. 6–8 2006. Modelica Association. URL `http://www.modelica.org/events/modelica2006/Proceedings/sessions/Session1a1.pdf`.

Francesco Casella, Tiemo Mathijssen, Piero Colonna, and Jos van Buijtenen. Dynamic modeling of organic rankine cycle power systems. *Journal of Engineering for Gas Turbines and Power*, 135(4):042310–1–12, 2013. doi:10.1115/1.4023120.

R. Franke, F. Casella, M. Otter, M. Sielemann, H. Elmqvist, S. E. Mattsson, and H. Olsson. Stream connectors - an extension of Modelica for device-oriented modeling of convective transport phenomena. In *Proceedings 7th International Modelica Conference*, pages 108–121, Como, Italy, Sep. 20–22 2009. The Modelica Association. ISBN 978-91-7393-513-5. doi:10.3384/ecp09430078. URL http://www.modelica.org/events/modelica2009/Proceedings/memorystick/pages/papers/0078/0078.pdf.

P. Fritzson. *Principles of Object Oriented Modeling and Simulation with Modelica 3.3*. Wiley IEEE Press, 2014.

S. E. Mattsson, H. Elmqvist, and M. Otter. Physical system modeling with Modelica. *Control Engineering Practice*, 6 (4):501–510, 1998.

M. Tiller. *Introduction to physical modelling with Modelica*. Kluwer, 2001.

Roberto Zanino, Roberto Bonifetto, Francesco Casella, and Laura Savoldi Richard. Validation of the 4C code against data from the HELIOS loop at CEA Grenoble. *Cryogenics*, 53:25–30, 2013. doi:10.1016/j.cryogenics.2012.04.010.

Method to Develop Functional Software for NPP APCS using Model-Oriented Approach in SimInTech

A.M. Shchekaturov, I.R. Kubenskiy, K.A. Timofeev, N.G. Chernetsov

Department of mathematical modeling
3V Services
Moscow, Russia
a.shchekaturov@3v-services.com

Abstract

The SimInTech environment developed in Russia enables using the end-to-end design of the algorithmic part of APCS for nuclear power plants (NPP), including the all-mode mathematical modeling of production process dynamics, debugging and optimizing control algorithms on an object model, generating functional software (FSW) as well as developing interfaces of the operator control panel. The article describes some of the main methods and approaches applied for the collective development of NPP APCS FSW. The implementation of the method during the development of APCS for Balokovo NPP-1 reactor compartment is presented as an example.

Keywords: model-oriented design, automated control system, algorithms, mathematical modeling, video frames

1 Introduction

A nuclear power plant, a ship with the NPP on board, and a submarine are all examples of the most complicated technical facilities produced by man. Designing a new NPP is always connected with a large scope of research and engineering activities, a group effort of specialist teams, the co-operation of several companies, and iterative calculations including those relating to optimization and search. In the process of NPP construction, external operating conditions should be considered along with the processes occurring inside the plant.

Automatic control systems (ACS) are an important NPP component that must comply with numerous, often contradictory, requirements. Control systems should feature a certain processing speed and a backup, ensure quality control and the required level of layered protection. Also, an option of manual control over production processes, an impactless mode-to-mode switch-over, the output of diagnostic information to the "upper" level, archiving signals with a specified stroke and others should be provided. It can be stated that all the specialists involved in the creation of APCS algorithms should know and understand thoroughly the physics of the phenomena, dynamics and technology of NPP processes.

As microprocessor technology develops, the rate of application of microprocessors in control circuits is increased; presently, we can discuss not only the control algorithms but also the functional software of ACS, which presents a non-trivial engineering problem. The complexity lies in the fact that not many process engineers can program, while programmers have a very little idea about technology. In general, process engineers articulate the ideas of algorithms in the language closest to them (including the functional block diagram language), whereas programmers interpret and implement these ideas in the way they have understood them, in a set of programs for specific equipment and the target operating system. At the same time, a large discrepancy may inadvertently appear between the software implementation conducted by the programmer and the initial ideas of the process engineer. It has to be eliminated at the stage of complex debugging and/or commissioning activities, which results in an increase in project costs and in a considerable loss of time.

In recent years, the model-oriented design (MOD) of control systems has been developed to a great degree (Voggenberger, 2005; Jakubowski, 2006; Dupleac, 2009) and has reached the level of a de facto standard to meet when generating FSW for ACS. Creating a mathematical model to describe the dynamics of a facility under design in a sufficient way forms the basis for the model-oriented approach. In the context of nuclear engineering, the MOD approach would be deemed optimal if it were not for technical complexities involved in its implementation, i.e. it is extremely difficult to create an all-mode model of NPP dynamics that would describe the dynamical characteristics of the facility in a satisfactory enough manner, it also represents single-purpose software that requires the application of design codes (thermo-hydraulic, neutronic, electrical, etc.).

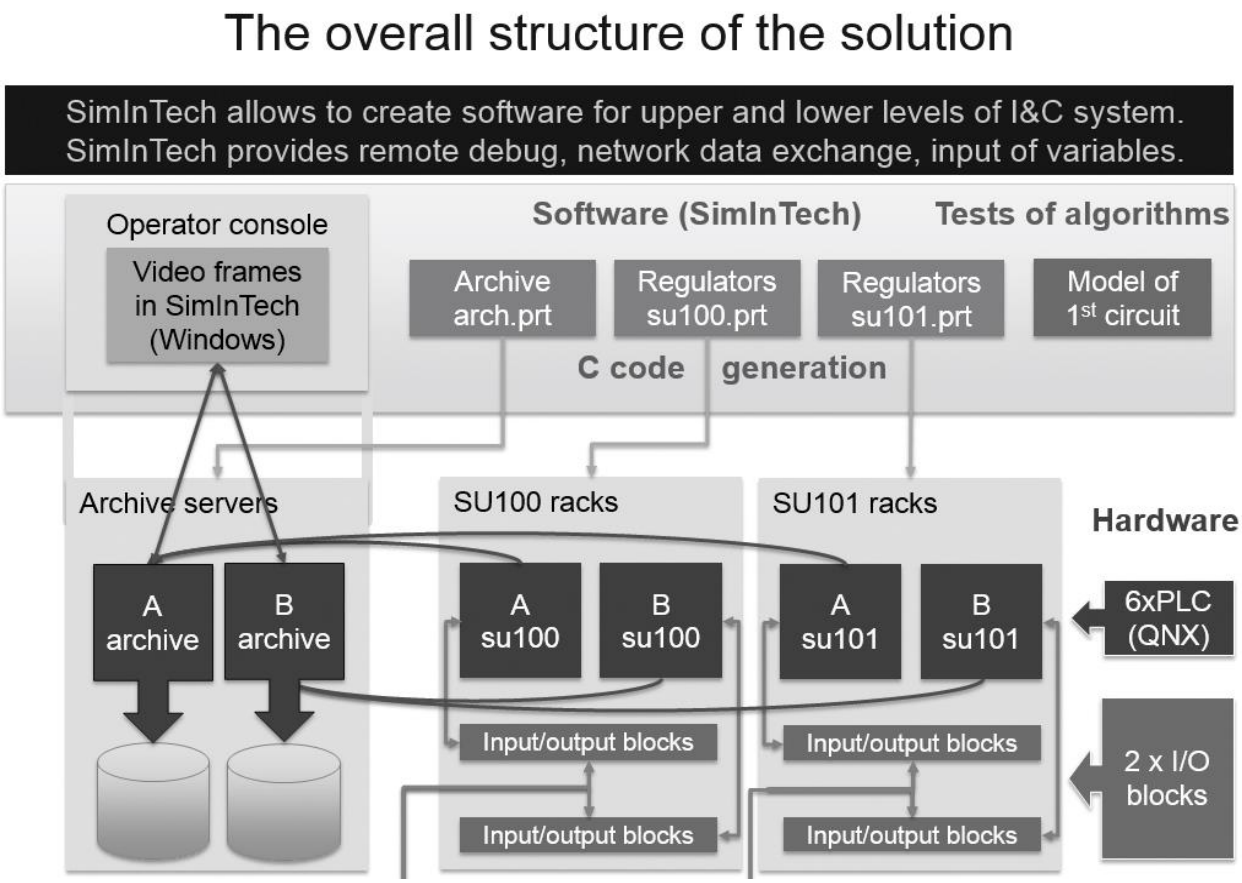


Figure 1. Hardware-Software Complex General View

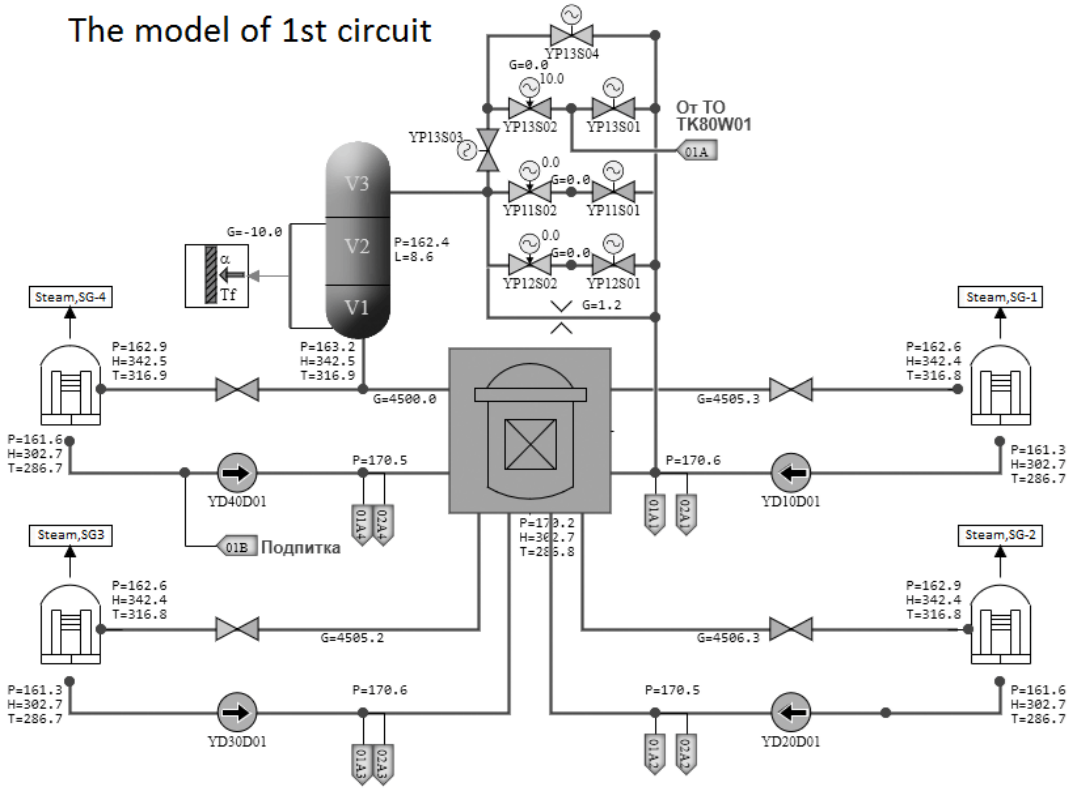


Figure 2. Design Diagram of the Primary Circuit Model

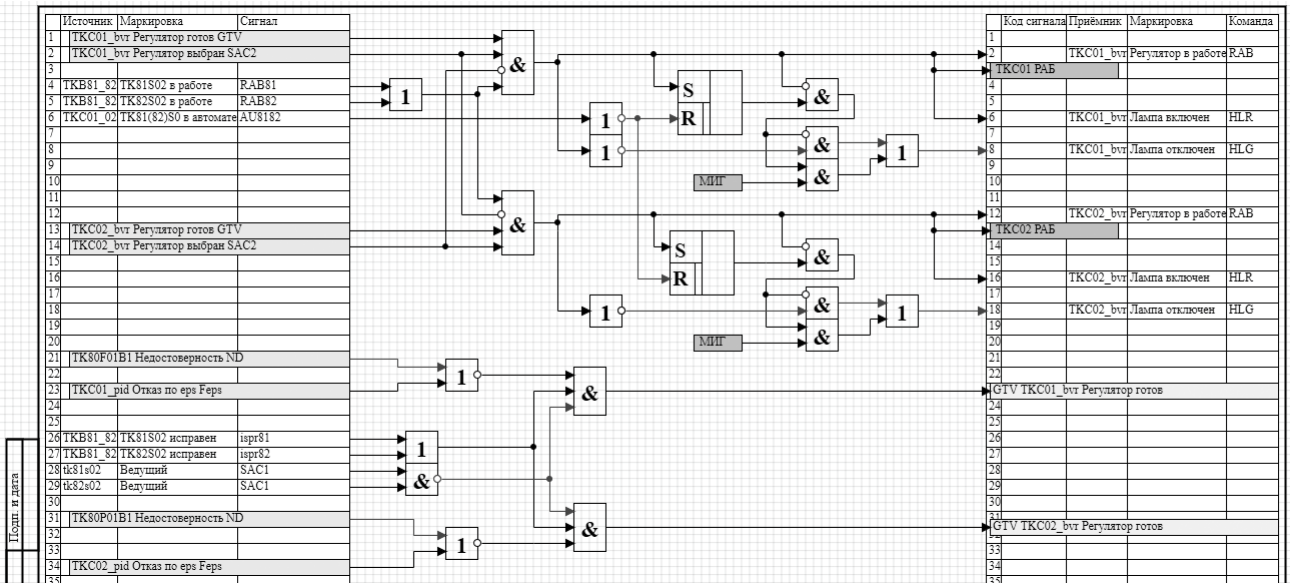


Figure 3. Example of a Functional Block-Diagram

Design codes are individual programs that have been developed for many years in design bureaus. Historically, they have not been developed for the purpose of integration and concurrent calculations with the use of any CAD of ACS algorithms. The SimInTech environment al-lows eliminating the problems relating to facility models and the integration of control algorithms thanks to its modular architecture, available interfaces to different design codes, and the relative simplicity of connection among new codes. The above fact has enabled the implementation of the "end-to-end" de-sign of the algorithmic part of APCS (Baum, 2012; Kozlov, 2005), i.e. control algorithms designed by technologists in the form of functional block-diagrams that are checked, debugged, and verified on an adequate model of facility dynamics are transmitted by the method of automatic generation of a code to the target sys-tem in the control equipment. As a result, prior to the stage of commissioning works, we obtain a product with verified FSW that is ready to be installed on the facility and requires virtually no adaptation or alteration.

Methods to Develop APCS FSW in SimInTech

The experience of applying the SimInTech environment to different pro-jects of mathematical modeling of NPP dynamics (Parshikov, 2013; Bezlepkin, 2013; Bolnov, 2014) and developing control algorithms evidentiates the main methods and techniques required for the development of ACS FSW. SimInTech authors believe that the sequence below can be of most use:

- stating the problem, identifying modeling and control boundaries, generating a list of equipment and signals, decomposing the facility into sub-systems, identifying boundary conditions; articulating agreement by the names of signals and variables;
- creating autonomous models of the physical processes occurring in the facility (models of processes of different physical nature, models of typical and unique pieces of equipment, models of actuators and devices, models of sensors, instruments);
- integrating the models into a complex model of dynamics of the facility;
- designing control algorithms for the facility in the form of functional block-diagrams, designing and modeling control panels and consoles;
- calculating concurrently and modeling facility dynamics and control system; testing and debugging the algorithms on the model, selecting and optimizing regulator coefficients;
- designing an analysis of the transient modes for normal operation and checking the operation of the automatic control algorithms;
- modeling design-basis and beyond-design-basis accidents as well as similar transient processes along with applying different equipment failures; adapting the algorithms;
- transmitting the control algorithms to the controllers, remotely debugging their implementation on the model;
- testing control equipment using testing devices which connect the mod-el as a simulator of the controlled facility by means of the inverse (digital-analog) transformation.

2 Applying SimInTech to Create Hardware-Software Complex of Automatic Regulation Facilities of Normal Operation Systems of Reactor Compartment

The hardware-software complex of the automatic regulation facilities of the normal operation systems of the reactor compartment provides the automatic regulation of the main process parameters of the first circuit: the coolant pressure in the primary circuit above the reactor core, the water level in the pressure compensator, makeup pump flow rates and other parameters. HSC also generates signals for complex measurements of such process parameters as the maximum mean coolant temperature in the primary circuit circulation loops, a difference of the primary circuit saturation temperature and the maximum temperature in the hot threads of loops, the material balance of the coolant for flushing-makeup of the primary circuit and others. Some rather strict requirements were specified for the system in terms of processing speed (algorithm stroke and parameter archiving – no longer than 20 ms), the number of registered and archived signals (over 15,000) and development periods (less than a year).

SimInTech has been used as the basis for developing a mathematical model of primary circuit dynamics (calculated by means of the TPP code), control and regulation algorithms, archiving server algorithms, a set of video frames and a system with analog input (see Figure 1). The code generator and executive environment included into SimInTech allowed the controllers to be programmed under the QNX real-time operating system, remote debugging of algorithm execution in the controllers to be carried out, and signals to be archived and displayed in video frames. The peculiarities of applying SimInTech to create HSC ARF RC are presented by implementing it in an integral design software environment for the upper and lower levels, including archiving system and system adjustment video frames:

- algorithm stroke – 10 ms;
- registration of 7,500 signals in the archiving system;
- back-up of the controllers and input/output channels;
- regulator parameter input system.

The project dealt with the challenge of modernizing the available automatic regulation facilities of the reactor compartment along with replacing the APCS equipment implemented on a basically different hardware platform. Available control algorithms presented as a text and block-diagram images were used as reference data. The sequence for creating automatic regulation facilities included:

1. Developing a mathematical model of the reactor compartment in a scope sufficient for debugging the algorithms and the primary adjustment of the control regulators (Figure 2).
2. Developing design diagrams of the algorithms and regulators (see Figure 3).
3. Testing the algorithms on the mathematical model of NPP RC, calculating all the required modes of operation (reactor start and shutdown, primary circuit heating-up and cooling-down, reactor power increase and reduction, nominal mode of operation).
4. Developing the algorithms for input-output signals digital processing – checking the authenticity of a signal, diagnosing the transducer off-scale sweep or breakout, the overrun of the tolerable speed of a signal change, filtration, signal conversion to a physical value and the required units of measurement.
5. Binding process signals to input-output blocks, generating the initial code automatically, programming the controllers, testing and debugging the execution of the algorithms on the controllers.
6. Testing ARF devices on the facility model.
7. Programming archiving servers, developing the algorithms of control channel "equalization" (setting the slave channel in compliance with the status of the master control channel in case of a restart of the controller).
8. Developing a set of video frames including the illumination of the in-put of regulator parameters in the process of operating the controller (Figure 4).

   This stage of commissioning works demonstrated an almost total absence of considerable errors– they were eliminated at the preliminary stages of developing and testing the algorithms.

3 Conclusions

Applying the SimInTech environment allowed the process engineers, programmers and test-operators of APCS to combine their efforts in an integral system, which led to a considerable (up to two-fold) reduction of the time required for developing a general software and hardware-software solution; it also minimized a number of errors in algorithm designing due to their elimination in the process of debugging on the facility model.

The application of this method in subsequent projects and in similar works by other teams has demonstrated the major practical advantage achieved due to the end-to-end design of control algorithms using the model-oriented approach. It is important to emphasize that the method developed on the basis of fully domestic tools provides as follows:

- exclusion of distortions in design solutions and their wrong interpretation by different groups of FSW designers (the minimization of the human factor);

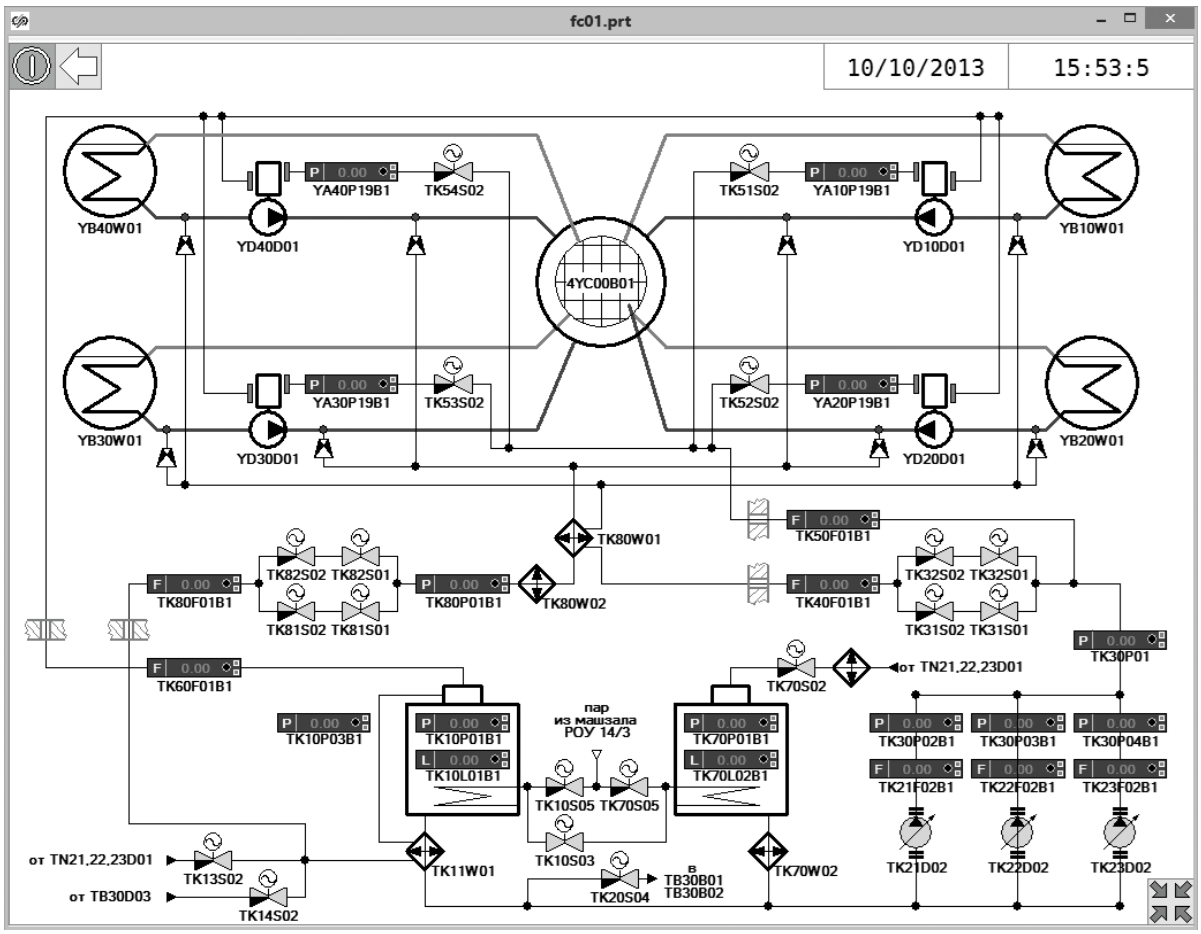


Figure 4. Example of Video Frame

- visualizing and debugging control system algorithms (reductions in the time for testing and debugging the instrument operation algorithm); a functional block-diagram is both a design model and a component part of FSW documentation;
- automated design testing of control system algorithms using mathematical models of the controlled object (reductions in labor costs allocated to programmers and process engineers);
- data systematization and structuring as per control system algorithms during the whole life cycle of the facility: from the development of the terms of reference up to the operational period.

References

F. I. Baum, O. S. Kozlov, I. A. Parshikov, V. N., Petukhov, K. A. Timofeev, A. M. Shchekaturov, SimInTech Software for Programming Control System Instruments. *Atomnaya Energiya (Atomic Energy) Magazine*, 113, July–December, 2012.

V. V. Bezlepkin, V. O. Kukhtevich, E. P. Obraztsov, Yu. A. Migrov, Hardware-Software Complex "Virtual NPP unit with VVER" (HSC "VPU") for Testing Design Solutions NPP-2006. *In the 8th International Scientific and Research Conference "Provision of Safety for NPP with VVER"*, OKB "Gidropress", Podolsk, Russia, May 28–31, 2013.

V. A. Bolnov, I. S. Zotov, S. A. Malkin, A. S. Ushatikov, Experience of Creation of Modeling Complex for NPP with RU BN-1200. *Atomny Proyekt (Atomic Project) Magazine*, 19, October 2014.

D. Dupleac, M. Mladin and I. Prisecaru, Generic CANDU 6 Plant Severe Accident Analysis Employing SCDAPSIM/RELAP5 Code, *Nuclear Engineering and Design*, 239: 2093-2103, 2009.

Z. Jakubowski, P. Dräger, W. Horche, W. Pointner, Development of Nuclear Plant Specific Analysis Simulators with ATLAS, *In proceedings of the 6th Conference on Nuclear Option in Countries with Small and Medium Electricity Grids* Dubrovnik, Croatia, 21-25 May 2006.

O. S. Kozlov, K. A. Timofeev, V. V. Khodakovsky. Software complex for research and development of dynamics and for designing of technical systems. *Information technology*, 9, 2005.

I. A. Parshikov, V. N., Petukhov, K. A. Timofeev, A. M. Shchekaturov, Modeling of Liquid-Metal Coolant Power Unit in SimInTech Software Complex. *University Scientific Magazine*, 5, 2013.

T. Voggenberger, D. Beraha and F.Cester, Nuclear Power Plant Simulation and Safety Analysis with ATLAS, *The 16th IASTED International Conference on Modelling and Simulation MS 2005*, May 2005, Cancun, Mexico

Object-Oriented Modeling with Rand Model Designer

Yu. B. Kolesov [1] Yu. B. Senichenkov[2]

[1]Mv Soft, Russia, ybk2@mail.ru

[2]Institute of Computer Science and Technology, Peter the Great St. Petersburg Polytechnic University, Russia, senyb@dcn.icc.spbstu.ru

Abstract

Rand Model Designer is a modern tool for modeling and simulation hierarchical multicomponent event-driven dynamical systems. It utilizes UML-based object-oriented Model Vision Language for designing dynamical and hybrid systems using modification of State Machines, and large-scale multicomponent systems: control systems with "inputs-outputs", "physical" systems with "contacts-flows", and novel variable structure component systems, particularly "agent" systems. This article provides a brief overview of the *«Object-Oriented Modeling with Rand Model Designer 7»* book contents (Kolesov et al., 2016), highlighting the differences between RMD and similar environments.

Keywords: object-oriented modeling, visual environment for modeling and simulation of event-driven complex dynamical systems, dynamical and hybrid systems, component "physical", "causal", and "agent"-based models, behavior of event-driven complex dynamical systems

1 Introduction

Object-oriented Modeling (OOM) is a modern technology of computer simulation (also referred to as computer modeling) widely used in scientific research, design of technical systems and analysis of business processes. OOM helps creating robust computer models in a faster and more cost efficient manner.

Computer modeling practically emerged at the same time with an appearance of first computers. At that point, computer models were created manually using programming languages like Fortran or even Assembly. For instance, one of the authors happened to develop his first computer model – imitator of the hardware (yet to be developed) on the test stand – with machine commands of BESM-4. Conversely, the level of abstraction of any designed system is significantly higher than the level of abstraction of any programming language. Thus, at the time, developer had to keep conformity between the model and its program implementation in his own mind, which, evidently, had led to numerous errors. Oftentimes, programmers used own program implementations of numeral methods to reproduce behavior of continuous systems, though this approach led to questionable results. Partially this issue had been addressed by the use of professional libraries (1980-1983): LINPACK, EISPACK, LAPACK (computational methods in linear algebra), ODEPACK (numerical methods for solving ordinary differential equations).

From 1950s, new high-level programming languages emerged, together with supporting software tools that created the executable computer models automatically. General Purpose Simulation System (GPSS) was one of the first modeling languages (Schriber, 1980). It was meant to be used as a simulator for queueing systems. This language is notable as it relied on object-oriented approach; however, this terminology was not used at that time. Transactions in the simulated system input were created as copies of the standard "Transaction" class, with different values of their attributes, and were destroyed after the execution.

Simulation programming language Simula-67 was released in late 1960s (Dahl et al., 1969). Simula introduced classes, objects, inheritance and polymorphism. Regrettably, this revolutionary project did not become widespread in the field of simulation practice, mainly due to limited computing power and high OOM use overheads of those days. At that time, emphasis was put on single-component isolated mathematic models, transcribed as large and unique equation systems, while the need of increasing speed and accuracy of computation dominated over other problems, such as structuring and modification of models, for instance. These circumstances led to the significant delay, measured in decades, in full-scale application of object-oriented approach in modeling. Conversely, Simula-type objects are re-implemented in a range of object-oriented programming languages, such as C++, Java and Object Pascal that are still in use today.

2 Matlab

MATLAB suite, probably one of the best-known universal simulation environments, was developed in late 1970s. It provides an opportunity to describe an isolated single-component dynamic system in a vector-matrix form, to call built-in solver for ordinary differential equations, as well as other solvers from professional systematic collections, and to visualize solutions in form of a time and phase diagrams. Despite

the fact that the models in MATLAB suite are described in algorithmic language, similar to Fortran, not in mathematical language, the suite development was a big step forward in the field, thus it justifiably gained the merited recognition.

Although MATLAB suite was primarily oriented toward creation of traditional mathematical models, an additional package added the functionality of component-based modeling. In 1992, this package acquired its present name - Simulink (Dyakonov, 2013). Graphical language of Simulink package has been used already at the time of analog machines and is familiar to developers of control systems – it allows building model of standard blocks, which have real physical equivalents in technical devices. Equations – often disliked by engineers – give way when graphical language of blocks becomes the main tool of model description. Simulink subsystem uses block language to describe models and exempts engineers from the time-consuming process of compiling system of equations corresponding with the whole model. The system of equations is then passed on to MATLAB suite solvers and the result of numeric solution is passed to different "visualizer".

Visual simulation technology of the Simulink suite comes down to an assembly of models from readily available "blocks" - elements of standard libraries with adjustable parameters. In essence, we are talking about OOM, as those "blocks" are principally objects of library classes. Well-designed library of basic classes allows creation of more complex classes by using available proven designs as elements. Structural complexity of new classes is disguised under multi-level hierarchical structure, thus complex classes on higher levels appear as basic elements. However, MATLAB's internal algorithmic language has to be used for any development or modification or of the abovementioned set of basic classes.

3 Unified Modeling Language

In late 1990s OOM, essentially, has had a major breakthrough in modeling practice, due to the introduction of the Unified Modeling Language (UML) (Booch et al., 2006) and the language for physical systems modeling Modelica (Modelica Association, 2012) reinforced with corresponding modeling software tools.

UML language is used to design specifications of complex software and hardware systems at the early stages of development. Its importance for discrete-event simulation lays foremost in canonization of concepts of object-oriented approach in software development, as it became de-facto standard of object-oriented approach. Moreover, these concepts were combined with exceptionally convenient and descriptive state diagrams (machines), invented in 1980's by D. Harel. UML language is easily extendable to simulate continuous-discrete (hybrid) systems. One indication of UML popularity could be seen in introduction of the State Flow subsystem, which supports state diagrams, to the Matlab suite.

4 Modelica language

Modelica language solved the problem of component simulation automation in the components with non-directional connections use. External variables of Simulink model components are referred to as "entrance" and "exit", which underlines the principal implication of the information transmission direction. However, in many application areas, there is no direction of external variables, for instance, the direction of currents and voltage symbols in electric circuits are rather conventional. Connection of external variables of different blocks in electrical circuits simply means the equality of the voltages and the equality of the sum of currents to zero. This fact fundamentally changes the method of constructing the final equations, which, in turn, leads to substantial problems. Modelica authors overcame these difficulties through introduction of components with non-directional connections, which in effect, broadened the range of application areas for computer simulation. Previously complex models were built using a limited set of basic components that could not be created in input simulating language, however today Modelica allows creating component libraries for electrical, hydraulic, mechanical, and other physical systems utilizing language's own capabilities. Nowadays, user could build any complex structure, where final equations for the structure are automatically generated, similarly to block models of Simulink. Moreover, Modelica classes could be inherited, defined and redefined.

5 Tools classification

Currently, there is a number of modeling tools, which support full OOM technology without narrowing focus to any of the application areas. These tools are divided in two groups:

1) tools focused on UML and its essential part - state machine formalism (hybrid automation);

2) tools focused on Modelica language (mechanisms supporting technology of "physical modeling").

First group includes environments like Ptolemy (University of California, Berkeley, USA) (Ptolemaeus, 2014), AnyLogic (The AnyLogic Company, Saint Petersburg) (Karpov, 2005) and Rand Model Designer (MVSTUDIUM Group, MvSoft and Peter the Great St. Petersburg Polytechnic University) (Kolesov et al., 2006; Kolesov et al., 2013). Additionally this group should also include the ever-developing Matlab that is now a complex system of components (MATLAB + Simulink + StateFlow + ToolBoxes). Matlab could be categorized as an object-oriented only with certain

provisions – OOM technology is fully supported only by the StateFlow subsystem within Matlab. However, this does not prevent Matlab from holding the leading role in modern applied simulation.

Second group includes environments developed within European project "Modelica", such as Dymola, Open Modelica, and MathModelica etc. For the purpose of this work, we consider OOM technology-based environments to be the most potential and thus we shall focus this discussion only on those.

Table 1. Object-Oriented approach.

Object-Oriented approach	
Simulink+Tollboxes	no
Modelica-based tools	yes
Ptolemy II	yes
RMD	yes

Modern simulating environment that claims to be universal has to allow development of the following models, in frameworks of the OOM technology:

- single-component continuous models
- single-component discrete-event models
- single-component hybrid models
- multi-component models with continuous, discrete or hybrid components and oriented connections (block models)
- multi-component models with continuous, discrete or hybrid components and non-oriented connections (physical models)
- multi-component models with variable composition of components and variable connection structure

Table 2. Universality. Definition.

Universality	
Dynamical and hybrid systems («isolated»)	Models with «input/output» components («causal»)
Models with «contact/flows» components («physical»)	Models with variable structure («agent-based»)

Table 3. Universality. Usage.

Universality	
Tollboxes (Simulink)	yes
Modelica-based tools	yes
Ptolemy II	no
RMD	yes

Undeniably, if the model has only two components and each one has own state machine with only to states – combined state machine will have for states, and every combined state will have an own corresponding system of equations to solve. Whereas, for components with non-directional connections, construction of aggregate system of equations cannot be reduced to a simple mechanical unification of component equations, typical for components with directed connections, and in general, requires a very complex analysis and transformation of equations. Matlab suite allows working with physical models, though only within frameworks of basic component sets.

Table 4. Compliance with UML.

Compliance with UML	
Simulink+Tollboxes	no
Modelica-based tools	partial
Ptolemy II	yes
RMD	yes

Modelica language supports "physical modeling", however it uses own concept of objects that does not match the one of UML language; it also uses special constructions to describe discrete events in hybrid models. Rather complex analysis of aggregate system of equations is conducted at the stage of model compilation; however, it could only be applied for a limited class of hybrid models, where the number of unknown variables and solved equations does not change with switching.

Modeling practice indicates that state machine of UML language is more convenient and graphical when it comes to the description of discrete-event and hybrid models than the description provided by the authors of Modelica language. Furthermore, numerous practically meaningful hybrid models are difficult to handle when using Modelica language. Besides, within the limits of Modelica's approach it is rather problematical to model systems with variable composition naturally.

6 Rand Model Designer

The Rand Model Designer tool attempts to combine the strengths of both directions: supporting the "physical modeling" suggested in Modelica language, while using object paradigm and states machine of UML language. Although, this design comes with an atonement - the need to implement part of the analysis of aggregate system of equations at the stage of model execution with every switch. It occurs that this analysis could be conducted by dint of "linear complexity" algorithms, and thus RMD – created industrial models, based on

components with non-directed connections, successfully work in real time. Differences of the input language of RMD environment – Manipulative Visual Language (MVL) language – and Modelica language are analyzed in (Martin-Villalba et al., 2014; Kolesov et al., 2014).

RMD environment features are briefly summarized below:

a. Elementary continuous behavior models

Continuous behavior may be represented in the form of implicit time dependencies, nonlinear algebraic equations (NAE), ordinary differential equations (ODE), and differential-algebraic equations (DAE). Equations are inputted by the user though the equation editor, similar to one in "mathematical" suites (Figure 1).

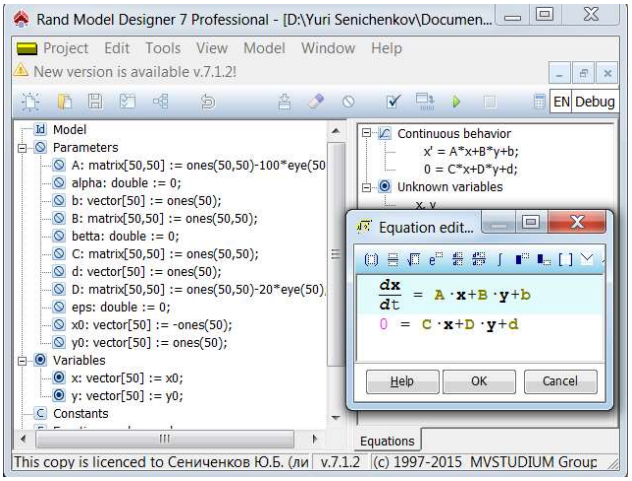


Figure 1. Example of continuous behavior.

Usage of both scalar and vector-matrix forms is possible. Alternatively, equations may be inputted in a block form, as it is done in the Simulink environment. SysLib database (Figure 2), which contains Simulink blocks, helps those users who are accustomed to a graphic description the continuous behavior.

b. Discrete and hybrid behavior models

Models with discrete and hybrid behavior are usually described using hybrid automata. Hybrid machines within RMD are UML state machines without parallel (orthogonal) activities (Figure 3).

Behavior charts feature “orthogonal” time. Many models with hybrid behavior require basing the selection of the next state on fairly complex calculations within the time gap. In particular, this may require modeling of additional dynamic systems, auxiliary to the core model. Thus, this simulation should be carried out in its own hybrid time, which is “orthogonal” to the principal time and is infinitely prompt – “instant”. Modeling of additional systems may be carried out in parallel, should the hardware capabilities suffice. Furthermore, RMD includes an option for keeping the specifications of any complex model as a class, and consequently consider abovementioned class’s behavior as "elementary" piecewise continuous state machine activity (Figure 4).

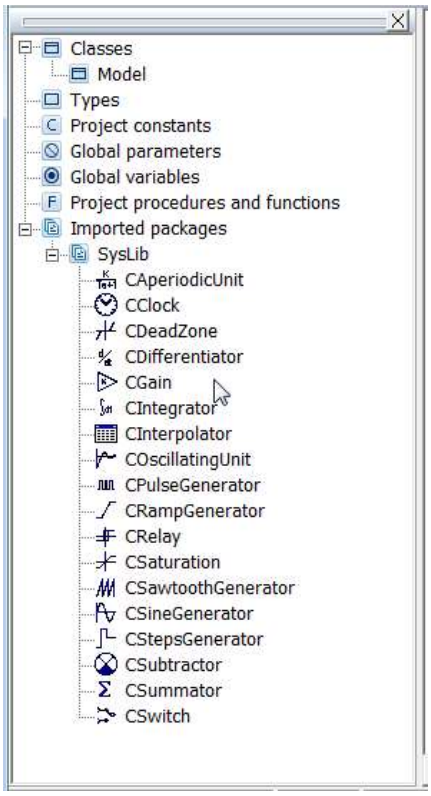


Figure 2. Library of “a la Simulink” blocks.

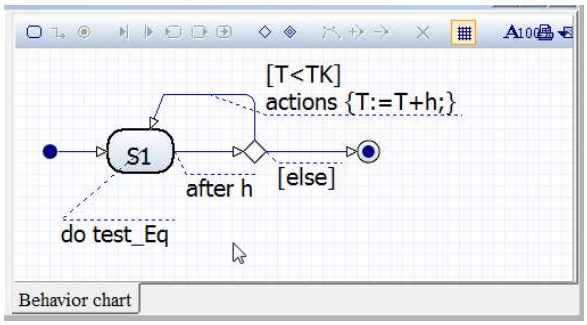


Figure 3. Behavior chart with UML notation.

c. Multi-component models with oriented components

Hierarchical multi-component models with components with the "input-output" and internal hybrid automata (with attributed continuous activities in form of equations or class instance activities, corresponding to complex subsystems models) demonstrate fairly complex behavior, which is best described by a component hybrid machine composition, with an extremely large number of states. It is not considered necessary to create this composition in an explicit form, as an appropriate state behavior could swiftly be created for the implementation of a specific event,

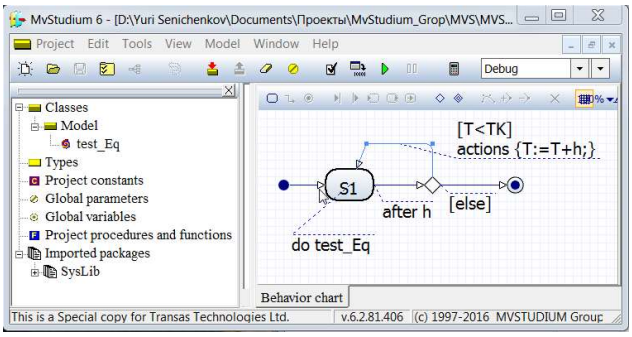


Figure 4. State activity as behavior of model converted into class.

d. Multi-component models with non-oriented components

Inheriting Modelica's technology for creating the multi-component model with oriented components (external variables such as "contact-flow") RMD introduces two significant additions:

- component hybrid automata may have continuous behavior, expressed as equation systems of any size and any type (NAE, ODE, DAE);
- equations, corresponding to component hybrid automata are created "on the fly" during the implementation, rather than at the compilation stage.

e. Multi-component models with variable structure

Construction of the variable structure models or models with "dynamic" components became possible owing to the introduced ability to create the equation for the whole model composition of component hybrid automata. Thus it is possible to create "agent-based models", as the component type (oriented or non-oriented) becomes irrelevant (Figures 5 and 6).

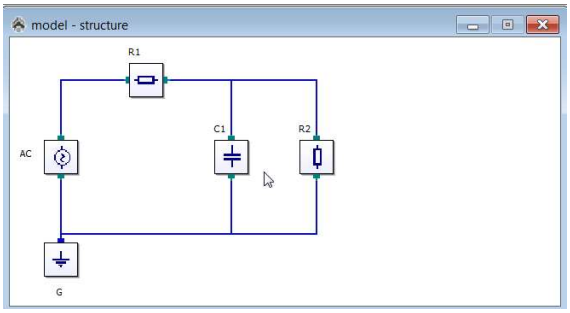


Figure 5. Initial structure of a component model.

Rand Model Designer version 7 allows creation of all the above mentioned types of models, based on the OOM technology. Trial version of RMD 7 is available at www.mvstudium.com.

Information about new book (Figure 7) you may find at http://www.labirint.ru/books/539673/ .

Tools of the first group – those that are based on UML – support all the aforementioned model types except for "physical models". This is due to the exponential growth of number of states in a state machine that corresponds to the whole model.

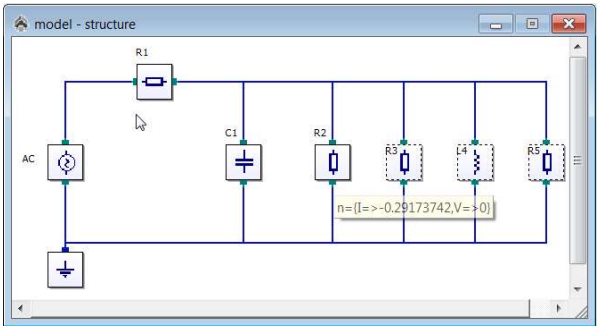


Figure 6. Structure of the model after the fifth event.

Figure 7. New book «Object-Oriented Modeling with Rand Model Designer 7».

References

G. Booch, I. Jacobson, and J. Rumbaugh. *UML 2.0*, Saint Petersburgh, Piter, 2006.

Claudius Ptolemaeus, Editor. *System Design, Modeling, and Simulation using Ptolemy II*, Ptolemy.org, 2014. – http://ptolemy.org/books/Systems.

O.J. Dahl, B. Murhaug, and K. Nigard. *Simula-67. Common Base Language.* - M: Mir, 1969.

V.Dyakonov. *Simulink: manual for self-tuition*. – M: DMK-Press, 2013.

Yu. Karpov. *Introduction in modeling and simulation with AnyLogic 5* – Saint Petersburgh: BHV-Petersburg 2005.

C. Martin-Villalba, A. Urquia, Y. Senichenkov, and Y. Kolesov. Two approaches to facilitate virtual lab implementation. *Computing in Science and Engineering*, 16(1): 79-86, 2014.

Modelica Association, Modelica® - *An Unified Object-Oriented Language for Systems Modeling*, Language Specification version 3.3. Modelica Association, 2012. - https://www.modelica.org (accessed in 2013).

Yu. Kolesov, Yu. Senichenkov, A. Urquia, and C. Martin-Villalba. Hybrid systems in Modelica and MvStudium. *Humanities and Science University Journal*, 8: 102-111, 2014.

Yu. Kolesov and Yu. Senichenkov. *Modeling of systems. Dynamical and hybrid systems*. Saint Petersburg, BHV, 2006

Yu. Kolesov and Yu. Senichenkov. *Mathematical modeling of hybrid dynamical systems*, Saint Petersburgh: Publishing of Polytechnic University, 2014.

Yu. Kolesov and Yu. Senichenkov. *Object-Oriented Modeling with Rand Model Designer 7.* Saint Petersburg, Prospect, 2016.

T. J. Schriber. *Modeling and simulation in GPSS.* — M: Mashinostroenie, 1980.

Rand Model Designer's Numerical Library

A. A. Isakov Yu. B. Senichenkov
Institute of Computer Science and Technology, Peter the Great St. Petersburg Polytechnic University, Russia
senyb@dcn.icc.spbstu.ru

Abstract

Numerical Libraries of visual simulation environments of complex dynamic systems differ from those of specialized collections of software implementations of numerical methods by means of the heuristic control program designed for automatic selection of numerical methods, accuracy control, and identification the specific features of systems of algebraic, ordinary differential and differential-algebraic equations on a local trajectories of the event-driven dynamic systems.

Keywords: dynamical and hybrid systems, component «physical», «causal», and «agent»-based models, behavior of event-driven complex dynamical systems, differential-algebraic equations, numerical methods for NAE, ODE, DAE

1 Introduction

Universal environments of visual simulation for complex dynamic systems (Simulink + Toolboxes, Modelica – based environments, Rand Model Designer, et cetera) are employed for construction, reproduction and behavior visualization of event-driven systems. Those come equipped with particular numerical libraries, which differ from general collections of professional numerical programs by means of heuristic control programs (also known as calling programs) for automatic selection of numerical methods.

Users of specialized collections, which are creating control programs on their own, usually deal with solution of a single system, or multiple systems, where system properties are pre-determined, or are determined in the modeling process. This allows users to pick out appropriate methods experimentally.

However visual environment users face exponentially large number of tasks generated by the state machine model, even when working with a simulation of a relatively small event-driven system. Systems of equations correspond to a particular state of a multi-component model and are built automatically; while a numerical method is routinely selected for those systems in the course of the process.

Reference article (Shampine et al., 2003) describes the software package MATLAB ODE Suite (The MathWorks) for numerical integration of event-driven dynamical systems. While referenced research works (Shampine et al., 2000; Shampine al., 1999) describe methods used for event location and for solving systems of differential-algebraic equations of high index.

Discussion on numerical problems in software environments which use Modelica language, are described in reference material (Fritzson, 2011; Sanfelice et al., 2010; Navarro-López, 2011; Senichenkov et al., 2013).

Rand Model Designer environment (Kolesov et al., 2013; Kolesov et al., 2014; Martin-Villalba et al., 2015; Kolesov et al., 2015; Isakov, 2015) contains a traditional library of software implementation of numerical methods (Fortran programming language), and a control module (C ++ programming language), which allows selection of specific solution methods for systems of equations on a given trajectory. Numerical Library and control module – hereinafter referred to as numerical library – are expressed in a form of dynamic link library (dll). This library provides the utmost rapid numerical solution for problems arising in the "release" mode, as well as debugging and tracing in the "debug" mode.

Open text software implementations of numerical methods (ODEPACK, RADAU, DDASPK ... - http://www.netlib.org) as well as solution algorithms of NAE, ODE, DAE presented in reference materials (Forsythe et al., 1977; Dongarra et al., 1979; Rice, 1981; George et al., 1981; Pissanetzky, 1984; Ascher et al., 1998; Shampine et al., 2003; Shampine et al., 2000; Shampine et al., 1999; Hairer et al., 1987; Hairer et al., 1991, 1996; Hindmarsh, 1983; Shampine et al., 1997), were used to create a RMD library. However, selected methods had to be modified and supplemented with original algorithms.

In addition to solution of equations on trajectories, it is also necessary to analyze the structure of the equations, to simplify the construction of the system, eliminating the variable "links", to determine state "switch-points" of event-driven systems, to ensure consistent initial conditions and to reduce index for differential-algebraic equations, as well as to determine the initial approximations for iterative methods.

Today visual simulation environments are preferred (Isakov, 2014):

- to support object-oriented modeling (OMM) technology
- to utilize modeling languages that allow user to simply describe most required tasks
- to comply with, not institutionalized, yet generally established, standards, such as, for instance Unified Modeling Language (UML) for discrete event-driven systems (Rumbauth et al., 2005).

Unification of the input language level only, as done for discrete event-driven systems (UML – http://www.uml.org), does not guarantee the same behavior in various environments of the same hybrid event-driven dynamic

system. Unification of the numerical libraries is required for optimal performance. It is desirable to achieve the same result as for the classified collections – when a given system, solved by numerical methods from different collections, has the same numerical solution for a predetermined comparison criterion.

In this article, we have tried to describe thoroughly the process of building, conversion, and numerical solution of systems of equations generated by the state machine of the hierarchical multicomponent event-driven model with variable structure within the visual modeling environment of the Rand Model Designer (RMD – www.mvstudium.com).

The Rand Model Designer (Kolesov et al., 2015) allows the model to be built in a form of an isolated dynamic or hybrid system, a hierarchical multi-component system of a permanent structure with components with the "input-output" external variables or components with "contact-flow", and model with variable (dynamical) structure.

2 Description of the complex dynamical systems behavior. Classical dynamical systems

Local continuous activity (continuous activity of hybrid automaton) could be expressed in form of:

- implicit dependencies on time
- linear and nonlinear algebraic equations (NAE)
- systems of ordinary differential equations in normal form or in form of equations not resolved with respect to derivatives (ODE)
- systems of differential-algebraic equations (DAE)

Each type of equations - NAE, ODE or DAE – has its own set of methods with mandatory program AUTO. Automatic mode is set by default, however user could replace it by any other method from the corresponding group.

Vector-matrix form of equations is preferred. It is "standard" form of equations for specific numerical method, which does not require further change when referring to a particular software implementation of the method.

However, especially in the early stages of design, user rarely thinks about the choice of method, and rather writes the equation in "free" form, which subsequently has to be converted into a "standard" form.

For example, even the simplest form of behavior description by way of scalar equations sequence with substitutions requires Equations Analyzer to regularize custom unknown variables and equations, to validate procedure and substitutions accuracy, as well as to determine “algebraic cycles”, i.e. hidden equations in those.

3 Consideration of solving systems structure

Solving of large-scale systems of nonlinear algebraic equations serves as the key task in the numerical simulation of complex dynamic systems. This task is important in itself; however, it becomes even more relevant in the context of ensuring consistent initial conditions, solving differential-algebraic equations, and implementing of the implicit methods for solving ordinary differential equation systems.

Within the RMD environment, Newton's method is utilized as a main method for solving systems of nonlinear algebraic equations. However, this method requires impeccable initial approximation. In the event of Newton's method failure, the task of finding solution is reduced to the task of minimizing a quadratic functional (Powell method).

The basic operation of the Newton’s method is solving systems of linear algebraic equations at each iteration. When using direct methods, it is important to pre-determine the matrix structure and call a corresponding modification of the method. Notably, RMD environment mostly relies on various modifications of the Gauss method for fully filled, band and sparse matrices

The implementation of the Gauss method for large sparse systems MA28 (Pissanetzky, 1984) provides the possibility of bringing the original system to block triangular form by means of rows and columns rearrangements. Whereas, bringing the original system to block-diagonal form, makes it possible to solve systems in parallel.

In case of block-triangular systems, only systems corresponding to diagonal blocks could have a natural solution. Consequently, the solution rate increases, but the question on optimal size of the diagonal block remains. Oftentimes, in case of dealing with small blocks, system solution overheads are inadequately high.

4 Standard description forms of continuous activity

In RMD equation and substitution differ syntactically. By default entry $x = f(c,k,t,y)$, where f - real-valued function (also known as real function), is considered as substitution, if variables $\{c,k\}$ on the right side are declared constant or parameters and value of variable y is determined at the time of calculation of the substitution. Here $t = \{Time, time\}$,

where ($Time$) – global, and ($time$) – local time.

Equations are written in a scalar form (1):

$$\begin{cases} z - q(c,k,t,z) = 0 \\ f(x,y,c,k,t) = 0 \\ g(x,y,c,k,t) = 0 \end{cases}, \qquad (1)$$

or vector-matrix form (2):

$$F(x,c,k,t) = 0, x, F \in \Re^n, \qquad (2)$$

and could be complemented by substitutions.

RMD gives a possibility for the user to "see" the system, which is being solved, in a designated window during the model execution. This is not only imperative for multi-component systems, with automatically built equations (based on component equations with consideration of their links), but also for the single-component systems, where final equation may differ from the original one.

User has to be notified about the Equation Analyzer operations: namely about allocation and regularization of required variables and equations. The process of allocation and regularization of variables and equations, as well as the one of determination of the equations structure is called structural analysis in RMD. It uses a bit-block system matrix, which indicates the occurrence of the unknown variables in the equation. The structural matrix for large systems takes a lot of memory space, while the structural analysis of the system takes a lot of time. This is especially true if system of equations is underdetermined and if conditions require finding a non-degenerate structural subsystem of maximum size. Oftentimes, at the early stages of building a new model, user needs Customer Support to eliminate underdetermined information.

Linear equations brought to the vector-matrix form

$$A(t)\cdot x(t)=b(t); \qquad (3)$$

are divided into the equations with fully filled, band and sparse matrix systems.

For systems with a fully-filled matrices, software implementation of the Gauss method with an assessment of the condition number of LINPACK package (Dongarra et al., 1979) (DGECO, DGESL) is used, for band matrices – subprograms (DGBCO, DGBSL). For systems with sparse matrices (Isakov, 2014) variation of the Gauss method (Pissanetzky, 1984) was realized, which provided the possibility of reducing a matrix to a block triangular form. It has been concluded that preliminary matrix reduction to block triangular form is justified for numerous technical tasks across various fields. Moreover oftentimes the final system turns out to be in a block-diagonal form, which makes it possible to solve systems in parallel. Taking into consideration specific properties of the diagonal blocks particular numerical methods may be employed when dealing with those. In particular, blocks where unit size equals one could be regarded as equations with a single variable, thus employing appropriate methods (Forsythe et al., 1977).

Nonlinear equations

$$F(x,c,k,t)=0 \qquad (4)$$

are solved with the Newton's method. In the RMD environment, in Newton's method, frequency of recalculation of the Jacobian matrix may vary, depending on the speed of convergence. Moreover method failure results in call in for the Powell method.

Ordinary differential equations with relation to the first derivative are recorded in scalar or vector form (5):

$$\frac{dx}{dt}=F(x,c,k,t),\quad x(0)=x_0;x,x_0,F\in\Re^n. \qquad (5)$$

Normal form, with relation to the second derivatives, is permitted (6)

$$\frac{d^2x}{dt^2}=F(x,c,k,t),\quad x(0)=x_0;\frac{dx}{dt}\Big|_{t=0}=V_0;\quad x,x_0,F\in\Re^n \qquad (6)$$

for the implementation of numerical methods that use it as the initial form (Hairer et al., 1987).

Differential-algebraic equations (7)

$$F(x,\frac{dx}{dt},c,k,t),\quad x(0)=x_0 \qquad (7)$$

are reduced to a semi-explicit form (8)

$$\begin{cases}\dfrac{dx}{dt}=x',x'(0)=x_0'\\ F(x,x',c,k,t)=0,x(0)=x_0\end{cases}, \qquad (8)$$

if those are not recorded in the following form (9)

$$\begin{cases}\dfrac{dx}{dt}=F(x,y,c,k,t),x(0)=x_0\\ G(x,y,c,k,t)=0;x,x_0,y,G,F\in\Re^n\end{cases}. \qquad (9)$$

It is possible to use explicit methods for the semi-explicit form (Hairer et al., 1987), resolving a nonlinear algebraic equations systems as required by a suitable method of NAE group.

Additional (in comparison with the solution of differential equations) operation is required when solving differential-algebraic equations – ensuring consistent initial conditions (utilizing suitable methods of NAE group).

The ability to successfully use the semi-explicit form depends on the properties of the Jacobi matrix. It is possible to use certain methods from the DAE group, for high index equation systems, if the system does not require conversion (Hairer et al., 1996). Alternatively it is possible to differentiate the equation to reduce the index.

Differentiation could be done symbolically, as in math suites, designated for symbolic computation (also known as

algebraic computation). Or alternatively, differentiation could be done numerically, as in visual modeling environments. Character differentiation block is planned to be implemented in Rand Model Designer in the near future. Meanwhile, differentiation is carried out by means of "linear differentiator".

Let $x = x(t)$ be the function, derivative of which is to be determined. Consider the following equation

$$\frac{d\tilde{x}'}{dt} = K \cdot (x - \tilde{x}'), \tilde{x}'(0) = \tilde{x}'_0. \quad (10)$$

For sufficiently large positive K solution to this additional equation tends to the specified function, while solution derivative tends to function derivative (if the derivative is "stable") (Emeljanov et al., 1997). Regrettably, even for the stable problems, error in the boundary layer may be excessive. However, experience has shown that for many models, even considerable errors in the boundary layer (arising from the unreliable initial conditions) are acceptable. It is important to keep in mind this source of potential errors.

5 Discrete dynamic and hybrid systems

RMD uses hybrid simulated time – continuous periods alternating with short time slots, where events are put in order (Kolesov et al., 2013; Kolesov et al., 2014). Discrete dynamic and hybrid systems are described by means of hybrid machines, equipped with a hybrid time.

To solve difference equations (11) a hybrid machine with an "empty" continuous activity and discrete actions is created

$$\begin{cases} Z_{k+1} := F(Z_k); k = 0,1,...; Z_0 = Z(0) \\ Z_k := Z_{k+1}; F, Z \in \Re^n \end{cases}. \quad (11)$$

The hybrid machine with continuous activities in the form of equations is essentially a generalization of classical dynamical systems, with all the inherent difficulties of numerical behavior reproduction - Zeno effect and sliding modes.

Hybrid machines within RMD are UML state machines without parallel (orthogonal) activities (Rumbauth et al., 2005), which relate to additional numerical tasks, namely determination of the "switch points", or event location, set by the operator "When", resulting in a state change.
Operator "When" may contain variables of any type, thus the main operational method for event localization is the method of bisection interval.

In case the event is associated with real variables and equality to a predetermined value, it is possible to formulate and solve the problem of finding a root on a given trajectory. For example, LSODAR program of ODEPACK – a collection of solvers for the initial value problem for ordinary differential equation systems (Hindmarsh, 1983). includes a root finding capability. However, it is difficult to determine the significance of this problem, due to the lack of data on the use frequency of operators "after", "signal", "when" complex system models.

6 Component systems with «inputs-outputs»

While components with "input-output" that have internal state machines produce exponentially large number of possible final systems, there are no major obstacles in creating those. Whatever way the components are connected, the structure of the component equations essentially does not change, if only left side variables of the substitutions that are defined as exit variables may become unknown variables of the new algebraic equations.
When using components with "input-output" the problem of differentiation of input variables is simplified, if they satisfy the differential equations within the components where they are formed. It is enough to pass the calculated value of the derivative to the right component.

7 Component systems with «contacts-Flows»

Components with "contact-flow" are utilized for "physical" simulation, where they generate an additional problem related to internal hybrid automata.

When using components with "input-output", if component equations for unknown variables are differential equations, then they remain differential equations in relation to the same variable also in resulting system for corresponding state in composition of component hybrid automata.

When using components with "contact-flow", if component equations for unknown variables are differential equations (differential variables) then in the resulting system they may turn into algebraic equations for composition of component hybrid automata. In result user's initial differential variables become algebraic. That is due to the concept of index of system of algebraic-differential equations (Hairer et al., 1996). Equation Analysers of visual simulation environments heave to be able to recognize the systems with high index and transform those in order to lower the index. This, in turn, requires differential equations. Some programs (Hairer et al., 1996) could solve systems of low index without reducing it, but recognition of such system remains a problem.

Resulting hybrid automaton of the whole model or the composition of component hybrid automata has a very large number of states even for relatively small models and requires significant efforts to form all the possible equations at the stage of compilation, in case of components with "contact-flow".

There are no restrictions on state equations of component automata in RMD, but the equations for realized composition states of component automata are being built at the time of model execution. It is an attempt to avoid building large number of resulting systems and practically to take into account the frequency of their appearance at a certain trajectory, while at the stage of execution only of system of realized states.

8 Component systems with variable structure

Processes of birth and death, queuing analysis, agent-based models all deal with dynamically changing number of objects (components), that are appearing and being destroyed leading to connections between them appearing and being destroyed. In this case the number of states of hybrid automaton of the whole model and the number of corresponding behaviours (systems of equations) becomes variable. Components may have any type of external variables e.g. “contact-flow”. For instance electric stations that have dynamic number of consumers, systems with reservation.

In such a case, creation of equations “on the go” becomes the only available means of creating equations. Fast algorithms of final system composition according to component equations and algorithms for connection equations for such models become one of the most important tasks.

9 Debugging

RMD has a debugger for conventional tracing debugging at the level of input language. The user cannot control the work of numerical methods on continuous sections by using debugger, but the user can use file “Tracing” that if required will hold fairly sufficient details about functioning of numeric algorithms, collected by using “print” operators provided by developers. It should be noted that the source texts of numerical libraries of visual simulation environments are not available to the users of visual simulation environments even if those are inevitable modifications of open specialized collections. This makes «trace.txt» file problematic to understand and to work with.

The more useful instruments are dynamic windows that are connected with the built and modified system of equations being solved at the moment: Final system, Structure Matrix, Jacobi’s Matrix, Eigenvalues of the Jacobi’s Matrix.

It is difficult to use all available information mainly because the user is creating own or using ready components with local equations, and receives information about automatically created resulting system, moreover large size systems are difficult to analyse.

Almost all the visual simulation environments face the problem of transformation of collected information into intuitive tips that help comprehending the behaviour properties of the whole model, indicate possible ways of model transformation in order to improve performance provide required precision and satisfy requirements for memory.

10 Timeline

The duration of continuous sections of trajectories is determined by the frequency of occurrence of events leading to a change in behaviour. Change of behaviour is usually accompanied by overhead costs for initialization of software implementation of numerical method. Selection of time progress step at each continuous section is determined by external circumstances and features of the problem being solved in combination with features of the selected method.

For example in case of solving nonlinear algebraic equations by iterative methods, internal step is determined only by external circumstances while taking in account the limitations associated with the choice of the initial approximation in form of solution from the previous step.

In case of differential equations, the internal step may considerably depend on the choice of the method. The rigidity of differential equations system can become an insurmountable obstacle for explicit methods. The problems with rapidly oscillating solutions are challenging for almost all methods if output information is needed with a large enough time step. Trying to increase step when solving rigid problems beyond minor boundary level contradicts with the need to save time for event localization.

In RMD the information about problem features on trajectory and costs of numerical solution is available in the “Timeline” tool. The user selects a desired constant advance step of model time and time interval to observe the model behaviour, and receives integral characteristics: time spent on execution of discrete actions, time spent on integration at continuous sectors, time spent on matching initial conditions when solving differential-algebraic equations.

11 Testing

Testing numerical libraries of the visual simulation environments also has specific characteristics.

Generally the libraries are sets of software implementations of numerical methods from open collections, modified according to the requirements of the environment: agreements on the acceptable forms of user equations, necessary information about behaviour visualization, information collected for processing the results of computational experiments with the model. The modifications lead to both errors and overhead costs. Both require testing that can be done by using existing collections of testing examples.

However, the information about overhead costs that occur as a result of model decomposition into components followed by aggregation of into single hybrid system of the whole model by the environment is considered to be of the most importance. At this point hybrid models of large size that can be built by hands and whose characteristics are known in advance are considered to be of the highest value. If decomposition of such models is possible by using components that are available in the environment, then it is possible to assess the overhead costs that occur when resulting hybrid automata are built by the environment. Besides the assessment of overhead costs it also becomes possible to compare the accuracy of solutions on trajectories. The authors are not aware of any test methods allowing verification of hybrid systems, besides a few minor attempts that manage to provide symbolic solution

(http://www.maplesoft.com/products/maplesim).

References

U.M Ascher and L.R. Petzold. *Computer methods for ordinary differential equations and differential-algebraic equations.* SIAM. Philadelphia, 1998.

J.J. Dongarra, C.B. Moler, J.R. Bunch, and G.W..Stewart. *LINPACK Users’ Guide*, Philadelphia, 1979.

S.V. Emeljanov and S. K. Korovin. *New types of feedbacks. M.: Science*, 1997.

G. Forsythe., M. Malcolm, and C. Moler. *Computer methods for mathematical computations*. Prentice-Hall, Englewood Cliffs, New Jersey, 1977.

P. Fritzson. *Introduction to Modeling and Simulation of Technical and Physical Systems with Modelica*, Wiley-IEEE Press, 2011.

A. George and W-H J. Liu. *Computer solution of large sparse positive define systems*. Prentice-Hall, Englewood Cliffs, Inc, New Jersey, 1981.

E Hairer, S. P. Norsett, and G. Wanner. *Solving ordinary differential equations I. Nonstiff Problems.* Springer-Verlag, 1987.

E. Hairer and G. Wanner. *Solving ordinary differential equations II. Stiff and Differential-algebraic Problems.* Springer-Verlag, 1996.

A. C. Hindmarsh. ODEPACK, A Systematized Collection of ODE Solvers, in *Scientific Computing*, R. S. Stepleman et al. (eds.), North-Holland, Amsterdam, IMACS Transactions on Scientific Computation, 1: 55-64, 1983.

A. A. Isakov, Yuri B. Kolesov, and Yuri B. Senichenkov. A new tool for visual modelling – Rand Model Designer 7. *IFAC-PapersOnLine* 48(1): 661-662, 2015.

A.A. Isakov OpenMVLShell visual environment. In Yu. B. Kolesov, and Yu. B. Senichenkov. *Mathematical modeling of hybrid dynamical systems.* St. Petersburg, Peter the Great Polytechnic University, 2014.

Yu. B. Kolesov and Yu. B. Senichenkov. Object-oriented modeling with Rand Model Designer. *The papers of annual scientific conference.* St. Petersburg, Peter the Great Polytechnic University, 6-12, 2015.

Yu. B. Kolesov and Yu. B. Senichenkov. *Mathematical modeling. Component technologies.* St. Petersburg, Peter the Great Polytechnic University, 2013

Yu. B. Kolesov and Yu. B. Senichenkov. *Mathematical modeling of hybrid dynamical systems.* St. Petersburg, Peter the Great Polytechnic University, 2014.

J. Lygeros, C. Tomlin, and S. Sastry. *Hybrid systems: Modeling, analysis and control.* http://inst.cs.berkeley.edu/~ee291e/sp09/handouts/book.pdf, 2008.

C. Martin-Villalba, A. Urquia, Y. Senichenkov, and Y. Kolesov. Two approaches to facilitate virtual lab implementation. *Computing in Science and Engineering* 16 (1): 78 – 86, 2014.

Eva M Navarro-López,and Rebekah Carter. Hybrid automata: an insight into the discrete abstraction of discontinuous systems. *Int. J. Syst. Sci.* 42(11): 1883-1898 2011.

S. Pissanetzky. *Sparse matrix technology.* Academic Press Inc. London, 1984.

J. R. Rice. *Matrix computations and mathematical software.* McGraw-Hill Book Company, New York, 1981.

J. Rumbauth, I. Jacobson and G. Booch. *The unified modeling language. Reference manual.* Second edition. Addison-Wesley, 2005.

Ricardo G. Sanfelice and Andrew R. Teel. Dynamical properties of hybrid systems simulators. *Automatica* 46(2): 239-248, 2010.

Y. Senichenkov. *Numerical modeling of hybrid systems. St. Petersburg,* Peter the Great Polytechnic University, 2004.

Y. Senichenkov, Y. Kolesov, and D. Inikhov. Rand Model Designer in Manufacturing Applications , *IFAC-PapersOnLine: Manufacturing Modelling, Management, and Control,* 7(1): 1572-1577, 2013.

L.F. Shampine, I. Gladwell, and S. Thompson. *Solving ODEs with Matlab.* Cambridge University Press, 2003.

L.F. Shampine and S. Thompson. Event location for ordinary differential equations. *Comput. Math. Appl*., 39:43-54, 2000.

L.F. Shampine, M.W. Reichelt, and J.A. Kierzenka. Solving index-1 DAEs in Matlab and Simulink. *SIAM review*, 41: 538-552, 1999.

L.F. Shampine and Mark W. Reichelt. The MATLAB ODE Suite. *SIAM Journal on Scientific Computing*, 18(1): 1-22, 1997.

Lena Wunderlich. *Analysis and numerical solution of structured and switched differential-algebraic systems.* Berlin: TU Berlin, Fakultät II, Mathematik und Naturwissenschaften (Diss.), 2008.

Adaptive Robust SVM-Based Classification Algorithms for Multi-Robot Systems using Sets of Weights

Lev V. Utkin Vladimir S. Zaborovsky Sergey G. Popov

Telematics Department, Peter the Great St.Petersburg Polytechnic University, Russia
lev.utkin@gmail.com,vlad@neva.ru,popovserge@spbstu.ru

Abstract

Three adaptive iterative minimax multi-robot system learning algorithms are proposed under condition that every observation obtained from robots is set-valued, i.e., it consists of several elements. The set-valued data are caused due to a fact that robots in the system provide different measurements in a single system observation. The first idea underlying the algorithms is to use sets of weights or imprecise weights of a special form for all elements of training data. The second idea is to apply the imprecise Dirichlet model for iterative updating the sets of weights depending on the classification accuracy and for assigning new weights to robots for improving classifiers. The simplest first algorithm is a modification of the SVM in order to take into account set-valued data. The second algorithm is the AdaBoost with the modified SVM under set-valued data. The third algorithm is the modification of the AdaBoost with updating imprecise weights of robots. The algorithms allow us to take into account the set-valued observations in the framework of the minimax decision strategy and to get optimal weights of robots to improve the classification accuracy of the trained multi-robot system.

Keywords: multi-robot system, SVM, classification, AdaBoost, set-valued observations, sets of weights

1 Introduction

Multi-Robot Systems (MRS) have drawn increasing attention last time due to several factors including the ability to perform complex tasks more efficiently compared to single-robot systems (Navarro and Matia, 2013; Tan and Zheng, 2013). In the MRS, robots are usually equipped with several sensors used to acquire as much information about the external world as possible. It is pointed out in (Pronobis et al., 2008) that each sensor may capture a different aspect of the environment, however, alternative interpretations of the information obtained by the same sensor can also be valuable. One of the important problems of the MRS learning is to integrate effectively multiple distributed sensor suites which may include GPS units, temperature sensors, altimeters, imaging systems, etc. (Cowley et al., 2004).

There are a lot of approaches for combining the multi-sensor information in robotics system learning (see, for example, (Du et al., 2012; Khamis et al., 2015; Ravet et al., 2013; Yuksel et al., 2012)) which mainly exploit various weighting schemes (in terms of probabilities or other measures) to differentiate the learning data sources and combine them in accordance with predefined rules taking into account the quality or reliability of data from different sensors. However, the most approaches assume that there is a large training set for learning and for assigning weights to sensors. This assumption may be violated in many applications, especially, during an initial learning phase when it is very difficult to estimate every robot or its sensors in order to apply the available weighted schemes.

Two main strategies of MRS learning can be marked out. The first strategy is when robots perform cooperative classification based on the same meta-classifier trained from training data obtained from sensors and from the feedback provided by a human teacher. This strategy may be useful during an initial learning phase when we do not know how different robots behave in their team and how reliable the sensor information provided by each robot. The second strategy is when each robot trains its own classifier (Di Caro et al., 2013), using the features extracted from a set of a locally available labeled examples, which correspond to the sensor information acquired from the robots' specific viewpoint and exploiting the experience from other robots.

The first strategy is studied in the present paper. We consider a case when it is difficult or just impossible to assign weights to separating sensors in order to combine their suites by using weighting schemes. The main difficulty in the joint use of training data from several sensors during the initial learning phase is that we cannot consider every sensor data as a separate training example. Let us consider for example the temperature sensors which provide with the environment temperature from a set of robots at some time moment. Every sensor provides with the information about the temperature of the same object approximately at the same time. Therefore, the set of temperature measurements in this case should be regarded as a single multi-viewed training example. Of course, we can use, for instance, some middle temperature for training. However, this combination rule says that all robots are identically reliable and accurate. This assumption is too strong in order to be valid in many applications. The initial learning phase is characterized by the lack of the

corresponding knowledge. Therefore, we propose a learning algorithm which takes into account the above peculiarities. It should be noted that only the initial learning phase may be available in some applications, and the learning algorithm proposed for this phase is entirely used in the MRS learning.

One of the most efficient and popular methods for the MRS classification learning is the support vector machine (SVM). Another efficient method is the AdaBoost proposed in (Freund and Schapire, 1997). Therefore, we propose a modification of the AdaBoost with SVMs of a special form as weak learners, which takes into account the fact that training data in the MRS are obtained from a set of unknown robots or their sensors. Moreover, we modify also SVMs such that imprecise judgements about robots can be incorporated into the SVM in order to improve the classification performance of the MRS. The main idea underlying the proposed modifications is the following. Training data from every robot are viewed as training examples, but their weights are replaced by some sets of weights such that every training example can be regarded as a set-valued data. Then we apply the robust minimax strategy in order to find an optimal decision function separating set-valued observations from different classes. The sets of weights are derived from the imprecise available information about robots. Moreover, we propose a double adaptive algorithm. The first adaptation is performed by the AdaBoost through changes of observation weights. The second adaptation is updating of the weight sets of set-valued observations in accordance with a number of correctly classified measurements from every robot at every iteration of the AdaBoost. The second adaptation is implemented by means of the imprecise Dirichlet model (Walley, 1996). In fact, we propose three algorithms for learning the MRS. The simplest one is just a modification of the SVM in order to take into account set-valued data. The second algorithm is the AdaBoost with the modified SVM under set-valued data. The third algorithm is the modification of the AdaBoost with updating imprecise weights of robots. The complexity of the algorithms do not differ from the complexity of the corresponding standard SVM and AdaBoost algorithms.

2 Formal Problem Statement and SVM

Suppose that we have observations or measurements from all sensors of T robots at every time moment k. After time moment n, we get the training set $S = \{(\mathbf{A}_1, y_1), ..., (\mathbf{A}_n, y_n)\}$, where $\mathbf{A}_k$ is a matrix having T rows $\mathbf{x}_1^{(k)}, ..., \mathbf{x}_T^{(k)}$, $k = 1, ..., m$, and m columns such that the row $\mathbf{x}_j^{(k)}$ is a vector of all measurements (features) obtained from the j-th robot. We assume that there are two classes (the binary classification) and $y_i \in \{-1, 1\}$. The learning aim is to construct an accurate classifier $c : \mathbb{R}^{m \cdot T} \to \{-1, 1\}$ that maximizes the probability that $c(\mathbf{A}_i) = y_i$ for $i = 1, ..., n$.

One of the ways for classification is to find a real valued separating function $f(\mathbf{x}, \mathbf{w}, b)$ having parameters $\mathbf{w}$ and b such that $\mathbf{w} = (w_1, ..., w_m) \in \mathbb{R}^m$ and $b \in \mathbb{R}$, for example, $f(\mathbf{x}, \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b$. Here $\langle \mathbf{w}, \mathbf{x} \rangle$ denotes the dot product of two vectors $\mathbf{w}$ and $\mathbf{x}$. We also denote $w = (\mathbf{w}, b)$. We assume for simplicity that, after the learning phase, every robot uses the separating function $f(\mathbf{x}, \mathbf{w}, b)$. Though we can also apply the function $f(\mathbf{A}, \mathbf{w}, b)$ which is defined for matrix $\mathbf{A}$ in the case of the inter-robot transfer learning.

One of the simplest ways for solving the classification problem is to replace every column of $\mathbf{A}_k$ by a number, for example, by the mean value of all elements of the column, and to apply the standard SVM.

In order to give a short description of the well-known SVM, we replace the set S by a set $S^* = \{(\mathbf{x}_1^*, y_1), ..., (\mathbf{x}_n^*, y_n)\}$. Here $\mathbf{x}_i^*$ is the vector of replaced values for every feature. Let ϕ be a feature map $\mathbb{R}^m \to G$ such that the data points are mapped into an alternative higher-dimensional feature space G. In other words, this is a map into an inner product space G such that the inner product in the image of ϕ can be computed by evaluating some simple kernel $K(\mathbf{x}_i^*, \mathbf{x}_j^*) = \left(\phi(\mathbf{x}_i^*), \phi(\mathbf{x}_j^*)\right)$ such as the Gaussian kernel. The SVM minimizes the empirical risk measure with a smoothness or penalty term $\langle \mathbf{w}, \mathbf{w} \rangle / 2$:

$$R(w) = \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle + C \sum_{i=1}^{n} l(y_i, \mathbf{x}_i^*, w). \tag{1}$$

Here C is the tuning "cost" parameter C which balances the trade-off between the empirical risk measure and the penalty term (Scholkopf and Smola, 2002); $l(y_i, \mathbf{x}_i^*, w)$ is the classification loss function. The so-called hinge loss function is used in SVM, i.e., $l(y, \mathbf{x}, w) = \max(0, 1 - y \cdot f(w, \phi(\mathbf{x})))$. Hence, the SVM classifier can be represented in the form of the following convex optimization problem with slack variables ξ_i, $i = 1, ..., n$:

$$\min_{\xi, w} R(w) = \min_{\xi, w} \left(\frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle + C \sum_{i=1}^{n} \xi_i \right) \tag{2}$$

subject to

$$\xi_i \geq 0, \;\; y_i \left(\langle w, \phi(\mathbf{x}_i^*) \rangle + b \right) \geq 1 - \xi_i, \; i = 1, ..., n. \tag{3}$$

The quantity $C\xi_i$ is the "penalty" for any data point $\mathbf{x}_i^*$ that either lies within the margin on the correct side of the hyperplane ($\xi_i \leq 1$) or on the wrong side of the hyperplane ($\xi_i > 1$).

Instead of minimizing the primary objective function (2) with constraints (3), we use a dual programming problem:

$$\max_{\alpha} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i^*, \mathbf{x}_j^*) \right) \tag{4}$$

subject to

$$\sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \le \alpha_i \le C,\ i = 1,...,n. \tag{5}$$

Here α_i, $i = 1,...,n$, are Lagrange multipliers or optimization variables in (4)-(5). After substituting the obtained solution into the expression for the decision function f, we get the "dual" separating function

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i^*, \mathbf{x}) + b. \tag{6}$$

The parameter b is defined by using support vectors $\mathbf{x}_i^*$ from the following equation

$$b = y_j - \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i^*, \mathbf{x}_j^*). \tag{7}$$

3 SVM by Set-Valued Training Data

The above approach for dealing with the training set S by replacing it with S^* cannot be used in the high-noise regime and by a rather small training set when robots provide scattered measurements. Therefore, in order to develop a robust classification procedure, we propose another approach for dealing with the training set S.

Let us consider a set of empirical expected risk measures $R(w)$ such that every measure from the set corresponds to a row, say $\mathbf{x}_j^{(k)}$, of the matrix $\mathbf{A}_k$. Then there exists an upper bound for $R(w)$, which is defined as

$$\overline{R}(w) = \max_{\mathbf{x}_i^{(k)} \in \mathbf{A}_k, k=1,...,n} \sum_{i=1}^{n} l(y_i, \mathbf{x}_i^{(k)}, w). \tag{8}$$

Here the expected risk is maximized over all $\mathbf{x}_i^{(k)}$ from $\mathbf{A}_k$, $k = 1,...,n$. The upper bound $\overline{R}(w)$ corresponds to the robust or pessimistic strategy in the sense that we select the "worst" elements $\mathbf{x}_0^{(k)}$ from $\mathbf{A}_k$.

Suppose that there are rows $\mathbf{x}_0^{(k)} \in \mathbf{A}_k$ for all $k = 1,...,n$, which provide the largest value of the expected risk. Then we can assign non-zero weights to the rows such that weights of other vectors $\mathbf{x}_i^{(k)} \neq \mathbf{x}_0^{(k)}$ become to be zero. This implies that the problem of maximization of the expected risk over rows of $\mathbf{A}_1,...,\mathbf{A}_n$ can be transformed to a problem of maximization of the expected risk over a set of weights. This transformation can be regarded as the uncertainty trick, i.e., we transform training data with the uncertainty of robot measurements to training data with the weight or probabilistic uncertainty.

Therefore, we extend the training set by rows $\mathbf{x}_i^{(k)}$ such that the extended training set has now $N = T \cdot n$ elements, but these elements have different weights. Let us denote a vector of new weights as $\pi = (\pi_1,...,\pi_N)$. Introduce also a set of indices $I_k = \{1 + (k-1)T,...,T + (k-1)T\}$. We only know about π that the sum of weights of all rows from $\mathbf{A}_k$ is $\sum_{i \in I_k} \pi_i = 1/n$ because every element of the initial training set has the weight or the probability $1/n$. This implies that the set $\mathscr{P}$ produced by all possible distributions π is convex, and there is an upper bound for $R(w)$, which is written as

$$\overline{R}(w) = \max_{\pi \in \mathscr{P}} \sum_{k=1}^{n} \sum_{i \in I_k} \pi_i l(y_i, \mathbf{x}_i^{(k)}, w). \tag{9}$$

It is important to point out that we did not simply extend the training set. By adding new elements to the training set, we change weights of the elements. At that, the weights of new elements are only partly known, and they belong to the set $\mathscr{P}$.

Now we can construct a modification of SVM taking into account the robust strategy, which is formulated as the following minimax optimization problem:

$$\min_w \overline{R}(w) = \min_w \max_{\pi \in \mathscr{P}} R(w). \tag{10}$$

Let us fix variables w and consider only a problem with variables $\pi \in \mathscr{P}$ by fixed w. The upper bound for $R(w)$ can be found by solving the optimization problem:

$$\overline{R}(w) = \max_{\pi \in \mathscr{P}} \sum_{k=1}^{n} \sum_{i \in I_k} \pi_i l(y_i, \mathbf{x}_i^{(k)}, w) \tag{11}$$

subject to

$$\sum_{i \in I_k} \pi_i = \frac{1}{n},\ k = 1,...,n,\ \sum_{k=1}^{n} \sum_{i \in I_k} \pi_i = 1. \tag{12}$$

The above constraints stem from the weights $1/n$ of initial training elements and from the sum of weights of all rows. It should be noted that the above optimization problem is linear and the following dual optimization problem can be written:

$$\overline{R}(w) = \min \left\{ c_0 + \frac{1}{n} \sum_{k=1}^{n} c_k \right\} \tag{13}$$

subject to $c_0,\ c_k \in \mathbb{R},\ k = 1,...,n,$

$$c_0 + \sum_{k=1}^{n} c_k \mathbf{1}(i \in I_k) \ge l(y_i, \mathbf{x}_i^{(k)}, w),\ i = 1,...,N. \tag{14}$$

Here c_0, c_k are new optimization variables; $\mathbf{1}(D)$ is the indicator function taking the value 1 if D is true. If we assume that sensor measurements are different for every training example, then the last constraints can be simplified as

$$c_0 + c_k \ge \max_{i \in I_k} l(y_i, \mathbf{x}_i^{(k)}, w). \tag{15}$$

Substituting the above constraint into the objective function, we get the upper expected risk

$$\overline{R}(w) = \min \left\{ \sum_{k=1}^{n} \max_{i \in I_k} l(y_i, \mathbf{x}_i^{(k)}, w) \right\}. \tag{16}$$

Substituting the hinge loss function into the objective function, adding the standard Tikhonov regularization term in order to restrict the class of admissible solutions and simplifying the problem, we get

$$\overline{R}(w) = \min\left(\frac{1}{2}\langle \mathbf{w}, \mathbf{w}\rangle + C \cdot \sum_{k=1}^{n} \xi_k\right) \tag{17}$$

subject to

$$\xi_k \geq 1 - y_k \cdot f(\phi(\mathbf{x}_i^{(k)}), w), \tag{18}$$

$$i \in I_k, \xi_k \geq 0,\ k = 1, ..., n. \tag{19}$$

The corresponding dual optimization problem (the Lagrangian) with variables α_i can be simply obtained

$$\max\left(-\frac{1}{2}\sum_{k=1}^{n}\sum_{t=1}^{n}\sum_{i\in I_k}\sum_{j\in I_t}\alpha_i\alpha_j y_k y_t K(\mathbf{x}_i^{(k)}, \mathbf{x}_j^{(k)}) + \sum_{k=1}^{n}\sum_{i\in I_k}\alpha_i\right) \tag{20}$$

subject to

$$\sum_{k=1}^{n}\sum_{i\in I_k}\alpha_i y_i = 0, \tag{21}$$

$$0 \leq \sum_{i\in I_k}\alpha_i \leq C,\ \alpha_i \geq 0,\ i \in I_k,\ k = 1, ..., n. \tag{22}$$

If we compare the above optimization problem with the standard SVM, then we can see that variables α_i are restricted in a different way (see constraints (22)).

If we assume that all points of intervals produced by using a grid are different, i.e., they are unique for every interval, then the objective function (20) can be rewritten as

$$\max\left(-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i^{(k)}, \mathbf{x}_j^{(k)}) + \sum_{i=1}^{N}\alpha_i\right). \tag{23}$$

The constraint (21) can be rewritten in the same way

$$\sum_{i=1}^{N}\alpha_i y_i = 0,\ \ \alpha_i \geq 0,\ i \in I_k, \tag{24}$$

$$0 \leq \sum_{i\in I_k}\alpha_i \leq C,\ k = 1, ..., n. \tag{25}$$

The separating function is of the form:

$$f(\mathbf{x}) = \sum_{k=1}^{n}\sum_{i\in I_k}\alpha_i y_i K(\mathbf{x}_i^*, \mathbf{x}) + b. \tag{26}$$

It can be seen from the above that the obtained SVM does not differ from the standard SVM with N training elements except for the last constraints where Lagrange multipliers are grouped in accordance with the robot information. When we have one robot, then $I_k = \{k\}, N = n$, and we get the standard SVM considered in the previous section.

4 A Modification of the AdaBoost

One of the efficient learning algorithms is the AdaBoost (Freund and Schapire, 1997). However, it is used for precise observations when the training set consists of point-valued examples. In order to improve the classification performance of the MRS, we propose a modification of the AdaBoost algorithm for the case of set-valued observations.

AdaBoost is a general purpose boosting algorithm that can be used in conjunction with many other learning algorithms to improve their performance via an iterative process. According to the AdaBoost, identical weights $h = (1/n, ..., 1/n)$ are initially assigned to all examples. In each iteration, the weights of all misclassified examples are increased while the weights of correctly classified examples are decreased (see Algorithm 1). As a consequence, the weak classifier is forced to focus on the difficult examples of the training set by performing additional iterations and creating more classifiers. Furthermore, a weight φ_t is assigned to every individual classifier. This weight measures the overall accuracy of the classifier. Higher weights are assigned to more accurate classifiers. The weight distribution $h(t)$ is updated using the rule shown in Algorithm 1. The effect of this updating rule is to increase weights of misclassified examples and to decrease weights of correctly classified examples. Thus, the weights tend to concentrate on "hard" examples. The final classifier c is a weighted majority vote of T weak classifiers. We assume below that classifier c_t is the weighted SVM.

Algorithm 1 The AdaBoost algorithm

Require: Q (number of iterations), S (training set)
Ensure: $c_t, \varphi_t,\ t = 1, ..., Q$
1: $t \leftarrow 1$; $h_i(1) \leftarrow 1/n$; $i = 1, ..., n$
2: **repeat**
3: Build classifier c_t using weights $h(t)$
4: $e(t) \leftarrow \sum_{i:c_t(x_i)\neq y_i} h_i(t)$
5: **if** $e(t) > 0.5$ **then**
6: $Q \leftarrow t - 1$
7: exit Loop
8: **end if**
9: $\varphi_t \leftarrow \frac{1}{2}\ln\left(\frac{1-e(t)}{e(t)}\right)$
10: $h_i(t+1) \leftarrow h_i(t) \cdot \exp\left(-\varphi_t y_i c_t(x_i)\right)$
11: $t \leftarrow t + 1$
12: **until** $t > Q$
13: $c(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{Q}\varphi_t c_t(\mathbf{x})\right)$

In order to modify the AdaBoost, we first provide a weighted version of the problem (20)-(22), namely, we suppose now the following condition for weights of observations:

$$\sum_{i\in I_k}\pi_i = h_k,\ k = 1, ..., n,\ \ \sum_{k=1}^{n} h_k = 1. \tag{27}$$

The dual problem for computing $\overline{R}(w)$ in this case is

$$\overline{R}(w) = \min\left(c_0 + \sum_{k=1}^{n} c_k h_k\right) \tag{28}$$

subject to c_0, $c_k \in \mathbb{R}$, $k = 1,...,n$, and (14).

It is simply to prove that the primal optimization problem for minimizing the upper expected risk is

$$\overline{R}(w) = \min\left(\frac{1}{2}\langle \mathbf{w}, \mathbf{w}\rangle + C \cdot \sum_{k=1}^{n} h_k \xi_k\right) \tag{29}$$

subject to (19).

The corresponding dual optimization problem with variables α_i differs from (20)-(22) only by constraints

$$0 \le \sum_{i \in I_k} \alpha_i \le h_k C,\ \alpha_i \ge 0,\ i \in I_k,\ k = 1,...,n, \tag{30}$$

where the upper bound for α_i is determined now by the weight h_k.

In order to use the AdaBoost, we define how to make decision about a class of a set-valued observation. A reasonable way is to apply one of the most popular strategy. According to the strategy, the set $\mathbf{A}_k$ belongs to a class y if at least a half of its elements $\mathbf{x}_j^{(k)}$ belong to the class y, i.e., there holds

$$y_k^* = \arg\max_{y \in \{-1,1\}} \sum_{i \in I_k} \mathbf{1}(c(\mathbf{x}_i^{(k)}) = y). \tag{31}$$

It is important to note that the proposed boosting algorithm deals with the extended training set consisting of N elements, but weights are modified only for set-valued observations, i.e., there are no restrictions for weights of elements from every set-valued observation except for the restriction (27).

5 Imprecise Updating Weights of Robots

So far we have considered the "worst" pessimistic case when we assumed almost total ignorance about the weight distribution over the robot measurements, i.e., we have assumed for the k-th local set of weights denoted as $\mathscr{P}_k$ the restriction $\sum_{i \in I_k} \pi_i = h_k$. Now we propose an adaptive algorithm for reducing the local sets of weights and for incorporating an additional information about weights of robots, which is defined by the classification errors at every iteration of learning. A main idea underlying the adaptive algorithm is to apply the imprecise Dirichlet model (IDM) proposed in (Walley, 1996) for updating local sets of weights at every step. This idea is close to an algorithm proposed in (Utkin, 2015), which uses the IDM in the AdaBoost in order to avoid a problem of overfitting.

Suppose that, after constructing a classifier on the basis of n set-valued observations at the t-th iteration of the boosting, we have $r_i^{(t)}$ correctly classified and $n - r_i^{(t)}$ misclassified measurements from the i-th robot, $i = 1,...,T$. This information allows us to update the local sets. It is reasonable to assume that if there are many misclassified measurements, then the set of weights $\mathscr{P}_k^{(t)}$ should be increased in order to make a robust decision. Moreover, it should be increased for robots whose measurements are misclassified.

In order to use the above information, we briefly consider the IDM. Let $U = \{u_1,...,u_T\}$ be a set of possible outcomes u_j. Assume the standard multinomial model: n observations are independently chosen from U with an identical probability distribution $\Pr\{u_j\} = p_j$ for $j = 1,...,T$, where each $p_j \ge 0$ and $\sum_{j=1}^{T} p_j = 1$. Then the IDM is defined in (Walley, 1996) as the set of all Dirichlet distributions over probabilities $p_1,...,p_T$ whose parameters are s and mean values $\mathbf{q} = (q_1,...,q_T)$ such that $\mathbf{q}$ belongs to the T-dimensional unit simplex denoted as $S(1,T)$. The hyperparameter s determines how quickly upper and lower probabilities of events converge as statistical data accumulate. Smaller values of s produce faster convergence and stronger conclusions, whereas large values of s produce more cautious inferences.

We propose to consider correctly classified measurements of the i-th robot as possible outcomes u_i. Then, according to Walley's IDM, we can write the bounds for the probability of $r_i^{(t)}$ correctly classified measurements as follows:

$$\frac{r_i^{(t)}}{D^{(t)} + s} \le p_i \le \frac{r_i^{(t)} + s}{D^{(t)} + s},\ i = 1,...,T, \tag{32}$$

where $D^{(t)} = \sum_{i=1}^{T} r_i^{(t)}$ is the total number of correctly classified measurements at the t-th iteration.

Note that we have $r_i^{(t)} = D^{(t)} = 0$ before iterations of the boosting. Hence, $0 \le p_i \le 1$. If we multiply the bounds on $h_k(t)$, then we get bounds for probabilities from the set $\mathscr{P}_k^{(t)}$, i.e., there holds

$$\frac{r_i^{(t)} h_k(t)}{D^{(t)} + s} \le \pi_{ki}^{(t)} \le \frac{\left(r_i^{(t)} + s\right) h_k(t)}{D^{(t)} + s},\ i = 1,...,T. \tag{33}$$

Here $\pi_{ki}^{(t)}$ is the weight of the i-th robot measurement in the k-th observation at the t-th iteration such that

$$\sum_{i=1}^{T} \pi_{ki}^{(t)} = h_k(t),\ k = 1,...,n. \tag{34}$$

Denote for simplicity

$$G_i = \frac{\left(r_i^{(t)} + s\right)}{D^{(t)} + s},\ F_i = \frac{r_i^{(t)}}{D^{(t)} + s},\ i = 1,...,T. \tag{35}$$

Then the dual optimization problem for computing $\overline{R}(w)$ at the t-th iteration is

$$\min\left(c_0 + \sum_{k=1}^{n} h_k(t)\left(c_k + \sum_{i=1}^{T}\left(g_{ki} G_i - d_{ki} F_i\right)\right)\right) \tag{36}$$

subject to $c_0,\ c_k \in \mathbb{R},\ g_{kj} \geq 0,\ d_{kj} \geq 0,\ j = 1,...,T,\ k = 1,...,n$, and

$$c_0 + \sum_{k=1}^{n} c_k \mathbf{1}(i \in I_k) + \sum_{k=1}^{n}\sum_{j=1}^{T} (g_{kj} - d_{kj})\mathbf{1}(I_k(j) = i) \geq l(y_i, \mathbf{x}_i^{(k)}, w),\ i = 1,...,N. \quad (37)$$

Here $I_k(j)$ is the j-th element of I_k. Let us write the Lagrangian by assuming that $l(y_i, \mathbf{x}_i^{(k)}, w)$ is the hinge loss function. It is of the form:

$$L = \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle + C \cdot c_0 + C\sum_{k=1}^{n} c_k h_k(t) - \sum_{k=1}^{n}\sum_{i=1}^{T} g_{ki}\lambda_i - \sum_{k=1}^{n}\sum_{i=1}^{T} d_{ki}\mu_i + C\sum_{k=1}^{n} h_k(t)\sum_{i=1}^{T}(g_{ki}G_i - d_{ki}F_i) - \sum_{k=1}^{n}\sum_{i=1}^{T}(\beta_{ki} + \alpha_{ki})(c_0 + c_k + g_{ki} - d_{ki}) - \sum_{k=1}^{n}\sum_{i=1}^{T}\alpha_{ki}(1 - y_k f(\phi(\mathbf{x}_i^{(k)}), w)). \quad (38)$$

Here $\alpha_{ki},\ \mu_i,\ \lambda_i,\ i = 1,...,T,\ k = 1,...,n$, are Lagrange multipliers. The saddle point can be found by setting the derivatives equal to zero. After simplifying, we obtain the following optimization problem:

$$\max\left(-\frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{i=1}^{T}\sum_{j=1}^{T}\alpha_{ki}\alpha_{lj}y_i y_j K(\mathbf{x}_i^{(k)}, \mathbf{x}_j^{(l)}) + \sum_{k=1}^{n}\sum_{i=1}^{T}\alpha_{ki}\right) \quad (39)$$

subject to

$$\sum_{k=1}^{n}\sum_{i=1}^{T}\alpha_{ki}y_k = 0, \quad (40)$$

$$\sum_{i=1}^{T}(\beta_{ki} + \alpha_{ki}) = Ch_k(t),\ k = 1,...,n, \quad (41)$$

$$CF_i \leq \sum_{k=1}^{n}(\beta_{ki} + \alpha_{ki}) \leq C\sum_{k=1}^{n} G_i. \quad (42)$$

Let us introduce new variables $\gamma_{ik} = (\beta_{ki} + \alpha_{ki})/C$. Then the above constraints except for the first one are rewritten as

$$F_i \leq \sum_{k=1}^{n}\gamma_{ki} \leq G_i,\ \sum_{i=1}^{T}\gamma_{ik} = h_k(t), \quad (43)$$

$$\alpha_{ki} \leq C\gamma_{ki},\ i = 1,...,T,\ k = 1,...,n. \quad (44)$$

In sum, we have derived a new optimization problem for computing the optimal values of α_{ki} and γ_{ik}. It is very interesting to see that constraints for γ_{ki} repeat the constraints for π_i. This is an important property of the obtained optimization problem. The optimal solution at the t-th iteration will be denoted as $\alpha_{ki}^{(t)}$.

The separating function at the t-th iteration is of the form:

$$f^{(t)}(\mathbf{x}) = \sum_{k=1}^{n}\sum_{i=1}^{T}\alpha_{ki}^{(t)}y_k K(\mathbf{x}_i^{(k)}, \mathbf{x}) + b^{(t)}. \quad (45)$$

After substituting the optimization problem (39)-(44) into the AdaBoost (Step 3 of Algorithm 1), we get the double adaptation. The first one is the updating of weights of observations. The second adaptation is the change of the sets of weights $\mathscr{P}_k^{(t)}$.

6 Conclusions

Three adaptive minimax SVM-based algorithms have been proposed in the paper. The first one can be regarded as a special case of the second algorithm. The second one can be also regarded as a special case of the third algorithm. The main peculiarity of the algorithms is that they use sets of weights instead of their precise values which are usually applied in many classification algorithm. The sets of weight are caused by the introduced transformation of uncertain set-valued training data from many robots to training data with the weight uncertainty in the form of these sets. The second peculiarity of one of the algorithms is that it is based on the use of Walley's imprecise Dirichlet model which allows us to reduce the sets of weights by repeating the classification procedure many times. An important property of the IDM is that it takes into account the prior total ignorance about weights before getting observations. The third peculiarity of the algorithms is their adaptivity. The sets of weights assigned to robots are adopted to classifiers. The fourth peculiarity is that the algorithms are robust because they use the minimax strategy for dealing with the weighted empirical risk measure under sets of weights.

It should be noted that the quadratic optimization problems which have to be solved for constructing the proposed classifiers are similar to the standard SVM optimization problems. Their difference is in additional linear constraints that, in fact, restrict sets of weights. In spite of the optimization problem simplicity, the standard software developed for the SVM in many packages, unfortunately, cannot be used. Therefore, the corresponding software has to be developed for implementing the proposed algorithms.

Acknowledgement

The reported study was partially supported by RFBR, research project No. 17-01-00118.

References

A. Cowley, H.-C. Hsu, and C.J. Taylor. Distributed sensor databases for multi-robot teams. In *Proceedings of the 2004 IEEE International Conference on Robotics and Automation, ICRA '04*, volume 1, pages 691–696, April 2004.

G.A. Di Caro, A. Giusti, J. Nagi, and L.M. Gambardella. A simple and efficient approach for cooperative incremental learning in robot swarms. In *Proceedings of the 16th International Conference on Advanced Robotics (ICAR-2103)*, pages 1–8, Nov 2013.

P. Du, J. Xia, W. Zhang, K. Tan, Y. Liu, and S. Liu. Multiple classifier system for remote sensing image classification: A review. *Sensors*, 12(4):4764–4792, 2012.

Y. Freund and R.E. Schapire. A decision theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1): 119–139, 1997.

A. Khamis, A. Hussein, and A. Elmogy. Multi-robot task allocation: A review of the state-of-the-art. In *Cooperative Robots and Sensor Networks*, volume 604, pages 31–51. Springer International Publishing, Cham, 2015.

I. Navarro and F. Matia. An introduction to swarm robotics. *ISRN Robotics*, 2013(Article ID 608164):1–10, 2013.

A. Pronobis, O.M. Mozos, and B. Caputo. SVM-based discriminative accumulation scheme for place recognition. In *Proceedings of the IEEE International Conference on Robotics and Automation, ICRA 2008*, pages 522– 529, May 2008.

A. Ravet, S. Lacroix, G. Hattenberger, and B. Vandeportaele. Learning to combine multi-sensor information for context dependent state estimation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5221–5226, Tokyo, Nov 2013. IEEE.

B. Scholkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, Massachusetts, 2002.

Y. Tan and Z.-Y. Zheng. Research advance in swarm robotics. *Defence Technology*, 9:18–39, 2013.

L.V. Utkin. The imprecise dirichlet model as a basis for a new boosting classification algorithm. *Neurocomputing*, 151(3):1374–1383, 2015. doi:10.1016/j.neucom.2014.10.053.

P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996. with discussion.

S.E. Yuksel, J.N. Wilson, and P.D. Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, 2012.

Network-Centric Control Methods for a Group of Cyber-Physical Objects

Vladimir Muliukha[1] Alexey Lukashin[2] Alexander Ilyashenko[2] Vladimir Zaborovsky[2]

[1] Almazov National Medical Research Centre, St.Petersburg, Russia, mulyukha_va@almazovcentre.ru
[2] Peter the Great St. Petersburg Polytechnic University, St.Petersburg, Russia, lukash@neva.ru, ilyashenko.alex@gmail.com, vlad@neva.ru

Abstract

In the paper we propose to use network-centric approach for a control task. Robots are described as cyber-physical objects that consist of two parts: mechatronic and informational. All cyber-physical objects are connected with each other using special multiprotocol nodes - devices that can route data between different types of computer networks (Ethernet, WiFi, 3G, LTE). Such network is described by hypergraph model, where central node is a hybrid cloud computer. While all robots are connected together, logical and computational tasks for cyber-physical objects are processed by this high performance node like in the central control system. Without a connection with central node robot switches into a multiagent mode.

Keywords: cyber-physical object, network-centric control, cloud robotics

1 Introduction

Modern trends in industrial technology include theoretical and applied research aimed at finding effective methods to control distributed dynamic systems. This research has particular importance for scientific foundations of field robotics that were greatly improved in recent years by the cyber-physical ideas. In this paper we provide constructive analysis focused on possibilities of application of cyber-physical approach to robots' motion planning and coordination to achieve control objectives in spatially and temporally undefined conditions. Proposed approach is based on multi-invariant actor-information representation and cooperative interaction of all components that describes a robotics system both on the physical (local or actor-based) and informational (knowledge or ontological-based) levels which are implemented using distributed resources of private cloud computing environment as IaaS and Hadoop. Actor-based representation (model) of each physical objects or artificial machine (robot) has specific attributes including name, data stack and parameters. The model describes environmental characteristics of mechanical, sensor, navigation and computer-communication components, the local interaction of which takes place onboard the robot and is needed to achieve declared control objectives. An informational model is proposed to represent common system characteristics, objective features, and robot as mobile carrier of specific operations. It is shown that system control requirements may be reduced to a "constraint satisfaction" problem. The decision of such problem is expressed by two sets of entities: a set of operations performed by robots of the group, and a set of messages that are generated by the informational model and delivered to each robots of the group using network infrastructure.

2 Cyber-Physics Approach

Complex engineering tasks concerning control for groups of mobile robots are not yet sufficiently developed. In our work, we use cyber-physical approach, which extends the range of engineering and physical methods for a design of complex technical objects by researching the informational aspects of communication and interaction between objects and with an external environment.

It is appropriate to consider control processes with cyber-physical perspective because of the necessity for spatio-temporal adaptation to changing goals and characteristics of the operational environment. Thus the priority task is to organize the reliable and high-performance system of information exchange between all entities involved in the realization of all requirements. Hereinafter, by cyber-physical object we mean an open system for the information exchange processes. Data in such system is transmitted through the computer networks, and its content characterizes the target requirements achieved through execution of physical and mechanical operations, energy being supplied by the internal resources of the object (Figure 1).

An example of a cyber-physical object is a mobile robot that does complex spatial movement, controlled by the content of the received information messages that have been generated by a human-operator or other robots that form a multi-purpose operation network. An ontological model of informational open cyber-physical object may be represented by different formalisms, such as a set of epistemic logic model

operations parameterized by data of local measurements or messages received from other robots via computer connection.

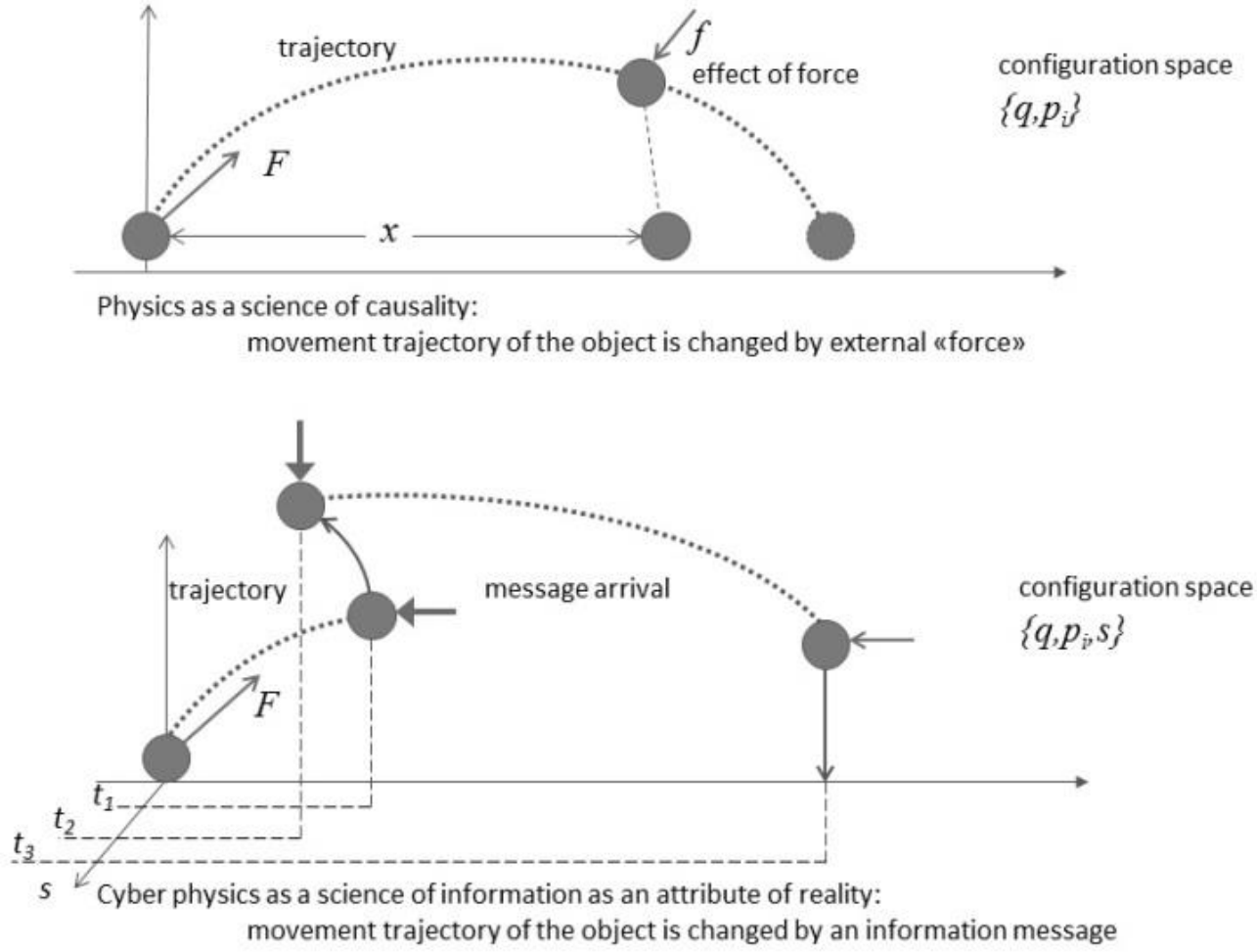


Figure 1. Physical and cyber physical motion

According to Figure 2, cyber-physical model of control system can be represented as a set of components, including following units:

- information about the characteristics of the environment (Observation),
- analysis of the parameters of the current state for the controlled object (Orientation),
- decision-making according to the formal purpose of functioning (Decision),

implementation of the actions that are required to achieve the goal (Action).

The interaction of these blocks using information exchange channels allows us to consider this network structure as a universal platform, which allows us to use various approaches, including the use of algorithms and feedback mechanisms or reconfiguration of the object's structure for the goal's restrictions entropy reduction or the reduction of the internal processes' dissipation.

Centralized solutions allow using universal means for the organization of information exchange to integrate different technologies for both observed and observable components of the controlled system. The parameters and the structure of such control system can quickly be adjusted according to the current information about the internal state of the object and the characteristics of the environment, which are in a form of digital data. These features open up the new prospects for the development of intelligent cyber physical systems that will become in the near future an integral part of the human environment in the information space of so-called "Internet of Things". According to the estimates (Zaborovsky et al., 2014), network-centric cyber-objects in the global information space of the Internet will fundamentally change the social and productive components of people's lives.

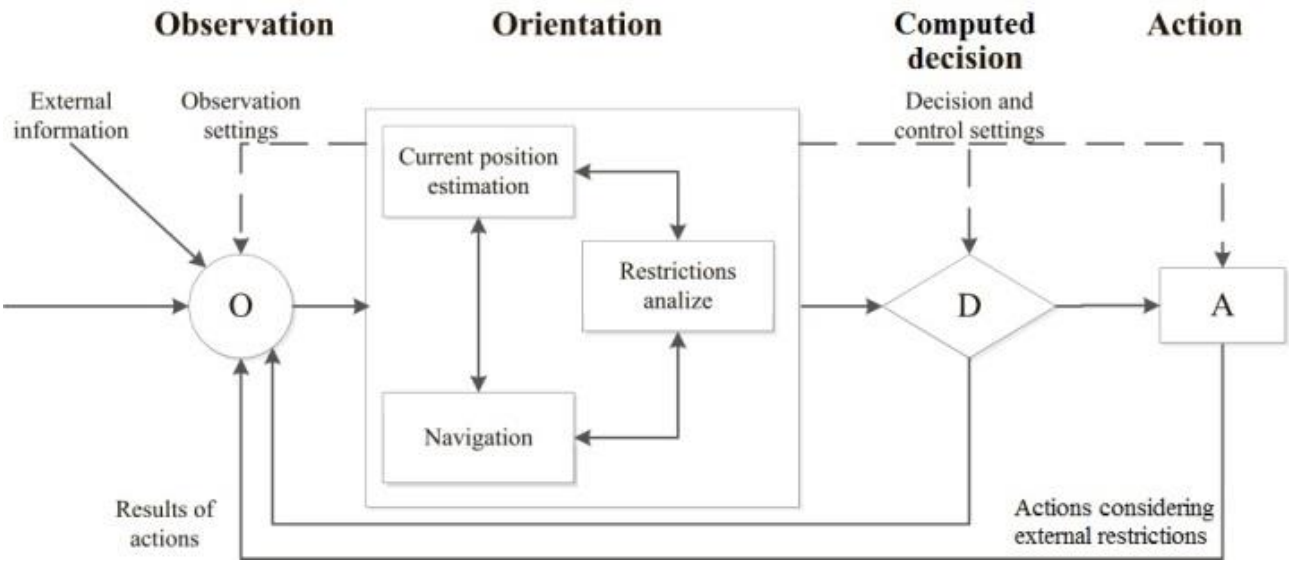


Figure 2. Cyber-physical interpretation of the John Boyd's OODA loop

The selection of cyber physical systems as a special class of designed objects is due to the necessity of integrating various components responsible for computing, communications and control processes («3C» – computation, communication, control). Therefore, the description of the processes in such systems is local and the change of its state can be described by the laws of physics, which are, in its most general form, a deterministic form of the laws of conservation of, for example, energy, mass, momentum, etc. The mathematical formalization of these laws allows us to determine computationally the motion parameters of the physical systems, using position data on the initial condition, the forces in the system and the properties of the external environment. Although the classical methodology of modern physics, based on abstraction of "closed system" is significantly modified by studying the mechanisms of dissipation in the so-called "open systems", such aspect of reality as the information is still not used to build the control models and to describe the properties of complex physical objects. In the modern world, where the influence of the Internet, supercomputers and global information systems on all aspects of the human activity becomes dominant, accounting an impact of information on physical objects cannot be ignored, for example, while realizing sustainability due to the information exchange processes. The use of cyber physical methods becomes especially important while studying the properties of systems, known as the "Internet of Things", in which robots, network cyber-objects and people interact with each other by sharing data in the single information space for the characterization of which are used such concepts as "integrity", "structure", "purposeful behavior", "feedback", "balance", "adaptability", etc.

The scientific bases for the control of such systems have become called Data Science. The term "Big Data" describes the process of integration technologies for digital data processing from the external physical or virtual environment, which are used to extract useful information for control purposes. However, the realization of the Data Science potential in robotics requires the creation of new methods for use of the information in control processes based on sending data in real time at the localization points of moving objects (the concept of "Data in motion").

3 Heterogeneous Platform for Cloud Robotics

The priority field of the development of science and technology are information and communications technologies and telematics services that are the foundation of modern infrastructure of the supercomputing engineering platform (Zaborovsky et al., 2014). The effectiveness of such a platform is ensured by cloud computing services that are implemented on the basis of hybrid supercomputing systems to effectively solve actual scientific and technical problems, including the following tasks: control of cyber-physical objects, predictive modeling and information security. The purpose of this paper is to solve the problem of the integration hybrid supercomputing resources in the control loop for a group of mobile robots, so called cyber-physical objects.

The base idea for the implementation of cloud technologies in the robotics is based on the paradigm of digital physics ("it from bit doctrine") (Wheeler, 1989), which implies that for robots the whole universe is measurable and computable. It means that if an object of the real "physical" world cannot be converted to the information from a sensor, for the robot and its control system such an object does not exist. So the key function of the robot's control system is to process

data received from sensors into the commands for actuators of the robot. For a number of situations, when such processing should not be performed in real time, it can be performed remotely in the cloud. The application of cloud for the solution for modern robotics tasks is called "cloud robotics" (Kehoe et al., 2015). The use of this approach reduces the processing load on each robot in the group and increases the efficiency of all cyber-physical objects by increasing the duration of battery life and reducing the redundancy of robots' characteristics for its tasks.

Researches in the field of cloud technologies in robotics are relevant and have been particularly active in recent years. For example, Rapyuta platform should be noted (the RoboEarth Cloud Engine). It allows robots to share knowledge, through a centralized knowledge base, avoiding duplication of information. Rapyuta is a cloud platform, in which the robots can create their own computing environments to perform intensive data processing. These created environments may be used with one or more robots simultaneously. This work is realized using the cloud based on modified OpenStack platform that is installed on heterogeneous high-performance computers containing SIMD and MIMD processors.

In this work, we offer the following solutions: a central server located on a secure heterogeneous cloud receives data from sensors of all available cyber-physical objects (mobile robots), processes the data and commands from the operator, and sends control commands to cyber-physical objects. The developed cloud software performs a decomposition of the complex target, which is set by operator, into the simple operation, transmits them to mobile robots, and checks their implementation, as well as provides synchronization of commands execution when it is needed. The use of the heterogeneous computing environment allows effectively realize a wide range of tasks, including processing of streaming data from cyber-physical objects (e.g., video or lidar data). An actor approach resolves the problem of dividing a whole system into several parallel streams allowing the horizontal and vertical scaling of the developed system. A functional hierarchical scheme of cloud computing environment is shown in Figure 3.

As part of this work a following concept was developed: each robot or cyber-physical object consists of mobile hardware and software part, so called "agent" and virtual "avatar" represented by a set of computational processes that are implemented in heterogeneous cloud computing environment, and are interacting with the agent using wireless high-speed networks. Such operating model of cyber-physical objects, in which onboard computing resources are supplemented by cloud computing environment's services, can solve a variety of control tasks, and provide a number of important benefits, including:

- advanced class of algorithms for control and operations planning because of integrating all available information resources from cyber-physical objects,
- ability to store and to structure large volume of sensory data,
- common data space and situational awareness of all agents in the system,
- possibility of rapid change in avatar algorithms (reboot avatar of cyber-physical object in real time), interaction between avatars through the high-speed data network in the cloud environment.

4 Practical Task: Constructing a Joint Map Task

We've considered one of resource-intensive tasks for the robotic group control. This task is to build and maintain the relevance of the joint map of the surrounding environment. The appearance of a controlled object is shown in Figure 4. An example of the proposed solution is shown in Figure 5.

Data from the lidar comes into onboard computer, which attaches to it a timestamp and the current position obtained from the navigation system. The resulting data are pre-filtered and sent to a high-performance cloud environment for further processing, synthesis and storage. Joint point cloud is formed in heterogeneous high-performance cloud environment using received lidar data with a timestamp and coordinates:

- an area of the existing joint point cloud is selected using coordinate information;
- common fragments in new and existing data are allocated;
- calculated translation and rotation matrix for the new point cloud;
- new point cloud is merged with the existing one and the resulting area of the joint point cloud is filtered.

Further, a global joint point cloud received from all mobile cyber-physical objects is vectorized. A map marked-up for passability is formed for each of robots in terms of its characteristics. It is possible to consider not only geographically but also telematic terrain characteristics, for example, a connection availability in a certain area.

All action of mobile cyber-physical objects are planned in high-performance cloud environment using marked-up maps. For example, when the operator sets the end point for the mobile robot, the route is laid within the permissible area defined by the marked-up map. During the movement of the robot the global joint point cloud and the map based on it are constantly updated. In this case, if the route goes beyond the

known area it will be corrected while new point cloud data will appear.

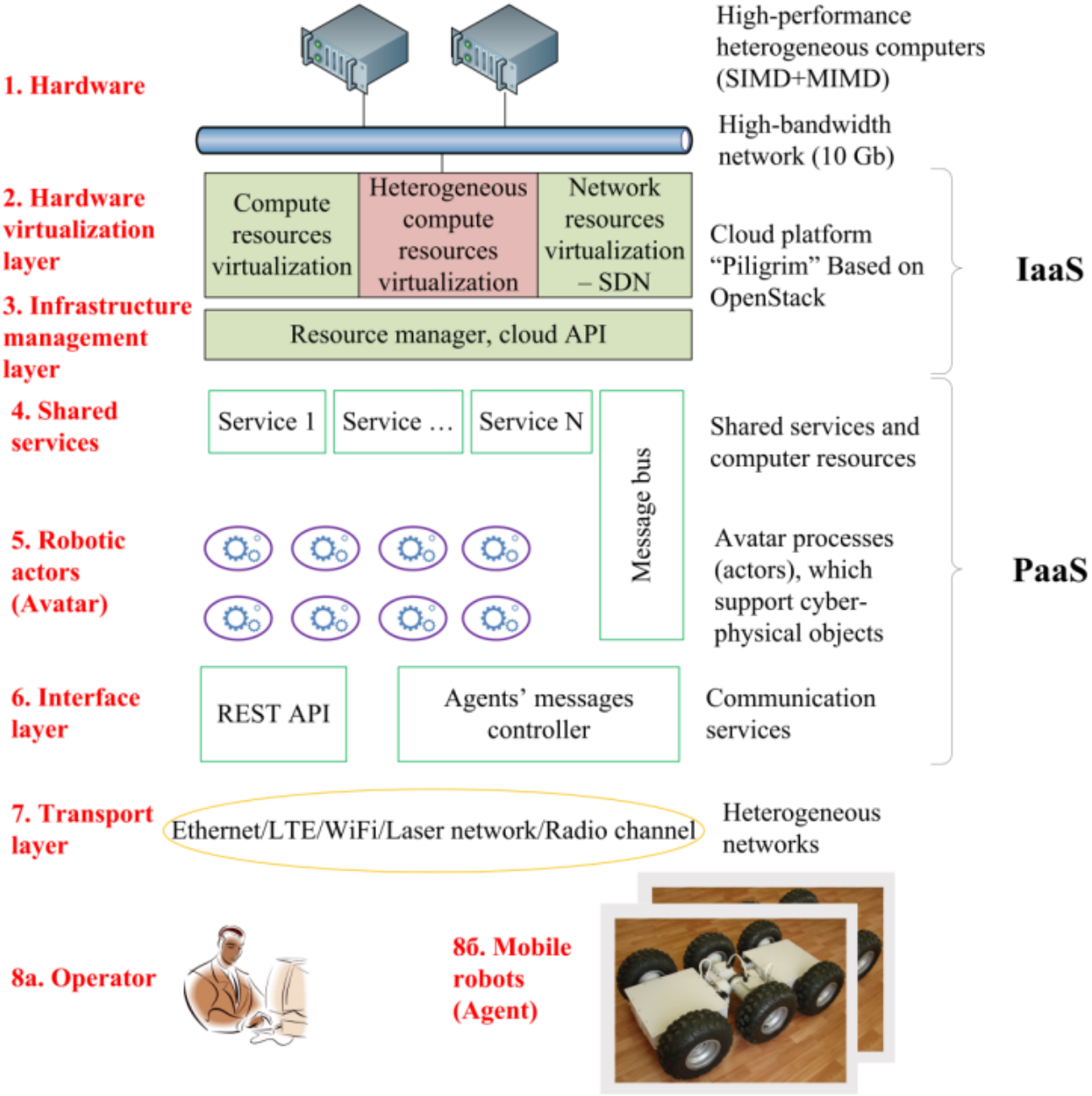


Figure 3. A functional hierarchical scheme of cloud computing environment

A fragment of generated route is transmitted to the onboard computer, which carries out the local control of the deviation between the current position and the desired trajectory. When the desired trajectory cannot be realized, the onboard computer requests new data from the high-performance computing environment.

In this work we use a point cloud library (PCL) and apply its existing methods. So one of the possible methods for point cloud vectorization is a mesh triangulation.

On the first step the point cloud is filtered to reduce the load on the compute nodes. Next, the triangulation is performed on the basis of data obtained. Depending on the degree of filtration it is possible to reduce the amount of data in more than ten times, while retaining sufficient accuracy of the model. An example of the proposed vectorization is shown in Figure 6.

5 Information Exchange in Network-Centric Control System

Main purpose of computer networks in network-centric control system is to transfer data between mobile robots (agents) and their virtual avatars. For working out effective transfer methods it is necessary to represent, what is the network, what processes proceed

in it and what influences its performance. To answer these questions it is necessary to develop a network model.

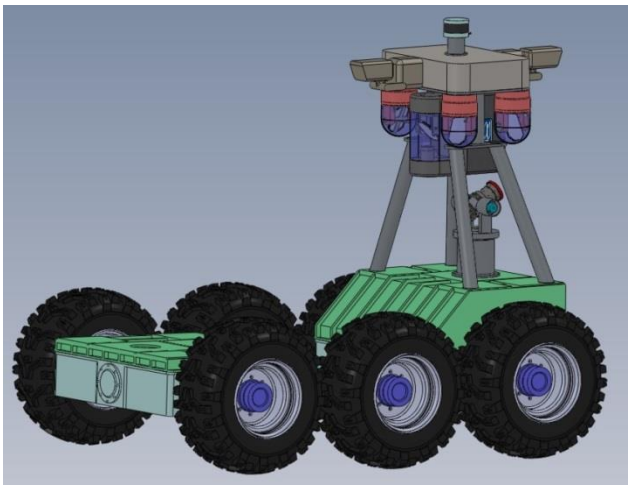


Figure 4. Appearance of controlled mobile robot

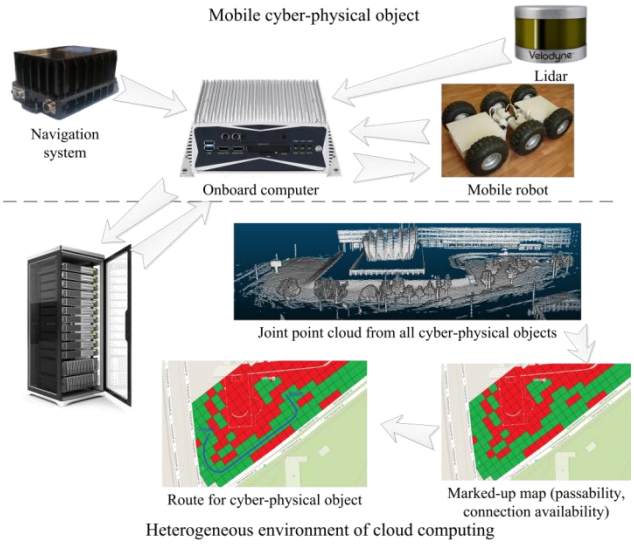


Figure 5. An example of using the cloud environment to control a group of robots: constructing joint map

Network-centric control system can be considered as data flows from various agents to their avatar applications. In our works (Ilyashenko et al., 2014; Ilyashenko et al., 2015; Muliukha et al., 2015) we consider the preemptive priority queueing system with two classes of packets. The packets of class 1 (2) arrive into the buffer according to the Poisson process with rate λ_1 (λ_2). The service time has the exponential distribution with the same rate μ for each class. The service times are independent of the arrival processes. The buffer has a finite size k $(1 < k < \infty)$ and it is shared by both types of customers. The absolute priority in service is given packets of the first class. Unlike typical priority queueing considered system is supplied by the randomized push-out mechanism. If the buffer is full, a new coming customer of class 1 can push out of the buffer a customer of class 2 with the probability α. Note that if $\alpha = 1$ we retrieve the standard non-randomized push-out.

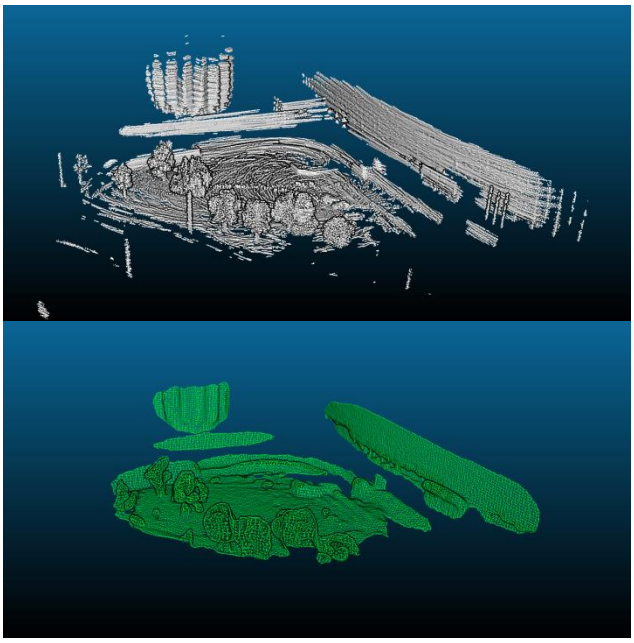


Figure 6. An example of the point cloud vectorization using PCL library

The summarized entering stream will be the elementary with intensity $\lambda = \lambda_1 + \lambda_2$. If we'll trace only the general number of packets in system, then simplified one-data-flow model would be $M/M/1/k$ type. The special modification of standard Kendel notation intended for priority queueing was proposed by G.P.Basharin. In the modified system the general structure of a label and sense of its separate positions remains however in each position the vectorial symbolic is used. There is an additional symbol f_i^j, where i specifies priority type (0 – without a priority, 1 – relative, 2 – absolute), and j specifies a type of the pushing out mechanism (0 – without pushing out, 2 – the determined pushing out).

We have analyzed received data and concluded that the computer network is modelled by means of queueing with the finite buffer size and the probability push-out mechanism in a combination with an absolute priority. The offered effective computing algorithm allows network engineers and designers to estimate possible variants of network traffic load. In overload networks, then $\lambda = \lambda_1 + \lambda_2 >> \mu$, the dependence of loss priority from variable α is not linear and even not monotonous function. In that case the network with an absolute priority gives a maximum of advantages to the most important types of packets. Our algorithm and network model is flexible and in comparison with a relative priority.

Queuing theory describes the process of information exchange in a network-centric robot group control system and ensures the required quality of service for each information flow between the agent and its virtual avatar. The proposed model for information exchange was used in practice in the framework of space experiments Kontur and Kontur-2 (Zaborovsky et al., 2015).

6 Conclusions

This paper proposes a network-centric approach to the description of the control system for a group of mobile robots.

Under this approach, each mobile robot is regarded as cyber-physical object, consisting of a mobile agent, and a virtual avatar that operates in a high performance heterogeneous cloud.

The use of network-centric approach allows to take an advantage of the combination of multi-agent and centralized control.

A necessary condition for the functioning of such a control system is the availability of reliable and high-speed communication channels between the agent and its avatar.

The reliability of these channels is ensured by using heterogeneous networks (LTE / Wi-Fi / 3G and etc.).

The required level of quality of service provided by prioritizing traffic based on the model of queuing theory.

Acknowledgment

This research was supported by RFBR grant № 15-29-07131 ofi_m.

References

A. Ilyashenko, O. Zayats, V. Muliukha, and L. Laboshin. Further Investigations of the Priority Queuing System with Preemptive Priority and Randomized Push-out Mechanism. *Lecture Notes in Computer Science* 8638: 433-443, 2014

A. Ilyashenko, O. Zayats, V. Muliukha, and A. Lukashin. Alternating Priorities Queueing System with Randomized Push-out Mechanism. *Lecture Notes in Computer Science* 9247: 436-445, 2015. doi: 10.1007/978-3-319-23126-6_38.

B. Kehoe, S. Patil, P. Abbeel, and K. Goldberg. A Survey of Research on Cloud Robotics and Automation. *IEEE Transactions on Automation Science and Engineering*, 12(2): 398-409, 2015

V. Muliukha, A. Ilyashenko, O. Zayats, and V. Zaborovsky. Preemptive Queueing System with Randomized Push-out Mechanism. *Communications in Nonlinear Science and Numerical Simulation* 21(1–3): 147–158, 2015

J.A. Wheeler. Information, Physics, Quantum: the Search for Links. In *Proceedings III International Symposium on Foundations of Quantum Mechanics*, Tokyo, pages 354-368, 1989

V. Zaborovsky, M. Guk, V. Muliukha, and A. Ilyashenko. Cyber-Physical Approach to the Network-Centric Robot Control Problems. *Lecture Notes in Computer Science* 8638: 619-629, 2014

V. Zaborovsky, V. Muliukha, and A. Ilyashenko. Cyber-Physical Approach in a Series of Space Experiments “Kontur”. *Lecture Notes in Computer Science* 9247: 745-758, 2015. doi: 10.1007/978-3-319-23126-6_69

Solving Stiff Systems of ODEs by Explicit Methods with Conformed Stability Domains

Anton E. Novikov[1] Mikhail V. Rybkov[2] Yury V. Shornikov[3] Lyudmila V. Knaub[4]

[1]Institute of Mathematics and Fundamental Informatics, Siberian Federal University, Russia, aenovikov@bk.ru
[2]Institute of Mathematics and Fundamental Informatics, Siberian Federal University, mixailrybkov@yandex.ru
[3]Automation and Computer Engineering Department, Novosibirsk State Technical University, shornikov@inbox.ru
[4]Institute of Mathematics and Fundamental Informatics, Siberian Federal University, Russia, lvknaub@yandex.ru

Abstract

The Cauchy problem for a stiff system of ODEs is considered. The explicit m-stage first order methods of the Runge-Kutta type are designed with stability domains of intermediate numerical schemes conformed with the stability domain of the basic scheme. Inequalities for accuracy and stability control are obtained. A numerical algorithm based on the first-order method and the five-stage fourth order Merson method is developed. The algorithm is aimed at solving large-scale systems of ODEs of moderate stiffness with low accuracy. It has been included in the library of solvers of the ISMA simulation environment. Numerical results showing growth of the efficiency are given.

Keywords: Runge-Kutta methods, accuracy and stability control, conformed stability domains, stiff problems

1 Introduction

Nowadays software for mathematical modelling and simulation is widely used for describing different processes in Chemical Kinetics, Electrotechnics and other applications. Models often are defined via either systems of ODEs or systems of PDEs. At that, systems of PDEs can be transformed to systems of ODEs applying discretization spatial derivatives. The greater the discretization step, the higher dimension of corresponding system of ODEs is. Furthermore, such problems are often stiff. This paper presents the algorithm of alternating order and step which is aimed at solving large-scale stiff problems with low accuracy. This algorithm has been included in the library of solvers (Novikov and Shornikov, 2012) of the ISMA simulation environment.

Consider the Cauchy problem for the stiff system of ODEs

$$y' = f(t, y), \quad y(t_0) = y_0, \quad t_0 \le t \le t_k, \tag{1}$$

where, y and f are sufficiently smooth real N-dimensional vector functions, t is an independent variable. Eigenvalues of its Jacobi matrix are pure real. It is well known that any initial value problem involving ODEs with higher derivatives can be reduced to this standard form. In (Hairer and Wanner, 1996; Novikov, 1997) for the solution of (1) the explicit Runge-Kutta methods

$$\begin{aligned} y_{n+1} &= y_n + \sum_{i=1}^{m} p_{mi} k_i, \\ k_i &= hf(t_n + \alpha_i h, y_n + \sum_{j=1}^{i-1} \beta_{ij} k_j), \end{aligned} \tag{2}$$

are presented, where k_i, $1 \le i \le m$, are stages of the method, α_i, p_{mi}, β_{ij}, $1 \le i \le m$, $1 \le j \le i$ - 1, are numerical coefficients, defining accuracy and stability properties of scheme (2). Methods of form (2) are rather efficient on solving non-stiff problems. However, from the numerical results of solving stiff problems with integration algorithms based on explicit formulas which choose stepsize according to the required accuracy it follows that on the settling region (where derivatives of a solution are low) there is plenty of declined solutions. This is a result of appearing instability of a numerical scheme.

Algorithms based on explicit methods with stability control of a numerical scheme can solve this problem. In this case previous errors are suppressed due to stability control, whereas new errors are low due to low values of solution derivatives. As a result, the practical accuracy is even greater than the accuracy, which is required. Further improvement of the efficiency can be reached on application of methods with conformed stability domains.

In (Novikov, 1997) the algorithm for obtaining coefficients of stability polynomials is presented. The use of these coefficients allows to design explicit Runge-Kutta m-stage methods for m equal up to 13 with defined form and size of a stability domain. It is also shown there that combining numerical formulas with different stability properties gives significant growth of performance. Transition from one numerical formula to another is performed according to stability criteria. At that, there is no explanation in (Novikov, 1997) how to choose coefficients β_{ij} which affect stability of intermediate (inner) numerical schemes and, finally, the efficiency of the integration algorithm.

The authors just noted that the stability of intermediate formulas can be achieved, if β_{ij} are chosen sufficiently small. Below the method for choice of coefficients β_{ij} is offered.

2 Numerical Schemes

For simplicity, here is considered the Cauchy problem for autonomous system of ODEs

$$y' = f(y), \quad y(t_0) = y_0, \quad t_0 \le t \le t_k, \tag{3}$$

but all the findings that are to obtained below stay true for non-autonomous problems, if the coefficients in (2) are defined by the formulas

$$\alpha_i = \sum_{j=1}^{i-1} \beta_{ij}, \quad 2 \le i \le m, \quad \alpha_1 = 0. \tag{4}$$

To solve problem (3) the Runge-Kutta methods of the following form

$$\begin{aligned} y_{n,i} &= y_n + \sum_{j=1}^{i} \beta_{i+1,j} k_j, 1 \le i \le m-1, \\ y_{n+1} &= y_n + \sum_{i=1}^{m} p_{mi} k_i, \end{aligned} \tag{5}$$

can be applied, where $k_i = hf(y_{n,i-1})$, $1 \le i \le m$, $y_{n,0} = y_n$, and $y_{n,i}$ are defined by formulas (2).

Introduce matrix B_m with elements b_{ij} (Novikov, 1997)

$$\begin{aligned} &b_{1i} = 1, \ 1 \le i \le m, \quad b_{ki} = 0, \ 2 \le k \le m, \ 1 \le i \le k-1, \\ &b_{ki} = \sum_{j=k-1}^{i-1} \beta_{ij} b_{k-1,j}, \ 2 \le k \le m, \ k \le i \le m, \end{aligned} \tag{6}$$

where β_{ij} are coefficients of scheme (2) or (5). It is to be used in the remainder of this paper.

Study stability on the linear scalar Dahlquist equation

$$y' = \lambda y, \quad y(0) = y_0, \quad t \ge 0, \tag{7}$$

with complex λ, $Re(\lambda) < 0$ (Dahlquist, 1963). Applying the second formula of (5) to (7), get

$$\begin{aligned} &y_{n+1} = Q_m(z) y_n, \ z = h\lambda, \ Q_m(z) = 1 + \sum_{i=1}^{m} c_{mi} z^i, \\ &c_{mi} = \sum_{j=i}^{m} b_{ij} p_{mj}, \ 1 \le i \le m. \end{aligned} \tag{8}$$

In the notations $C_m = (c_{m1}, \ldots, c_{mm})^T$ and $P_m = (p_{m1}, \ldots, p_{mm})^T$, the third relation of (8) can be written in the form

$$B_m P_m = C_m, \tag{9}$$

where the elements of matrix B_m are defined by relations (6). For intermediate numerical schemes (4) we have

$$\begin{aligned} y_{n,k} &= Q_k(z) y_n, \ Q_k(z) = 1 + \sum_{i=1}^{k} c_{ki} z^i, \\ c_{ki} &= \sum_{j=i}^{k} b_{ij} \beta_{k+1,j}, \ 1 \le k \le m-1. \end{aligned} \tag{10}$$

On $\beta_k = (\beta_{k+1,1}, \ldots, \beta_{k+1,k})^T$ and $c_k = (c_{k1}, \ldots, c_{kk})^T$ coefficients β_{ij} of numerical schemes (5) and the coefficients in the corresponding stability polynomials satisfy the equation

$$B_k \beta_k = c_k, \quad 1 \le k \le m-1. \tag{11}$$

From the comparison between (6) and (10) it follows that $b_{ki} = c_{i-1,k-1}$, i.e. the elements of (k+1)-th column of matrix B_m equal to coefficients of stability polynomial $Q_k(z)$. Hence, if the coefficients of stability polynomials of basic and intermediate numerical schemes are defined, then the coefficients of methods (5) are unambiguously determined from linear systems (9) and (11) with upper triangular matrices B_i, $1 \le i \le m$.

Expansions of the exact and approximate solutions in the Taylor series in powers of h have the form

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + hf + 0.5h^2 f'f + O(h^3), \\ y_{n+1} &= y_n + (\sum_{j=1}^{m} b_{1j} p_{mj}) hf + \\ &\quad (\sum_{j=2}^{m} b_{2j} p_{mj}) h^2 f'_n f_n + O(h^3), \end{aligned} \tag{12}$$

where the elementary differentials are computed on exact $y(t_n)$ and approximate y_n solutions, respectively. Comparison between relations (12) under assumption that $y(t_n) = y_n$, shows that numerical formula (5) has the first order of accuracy, if $\sum_{j=1}^{m} b_{1j} p_{mj} = 1$. Hence, to design m-stage methods of the first accuracy order, it is necessary to set $c_{m1} = 1$ in linear system (9).

3 Conformation of Stability Domains

Assume that the coefficients of the stability polynomials

$$Q_k(z) = 1 + \sum_{i=1}^{k} c_{ki} z^i, \quad 1 \le k \le m. \tag{13}$$

are defined. Using approach from (Novikov and Rybkov, 2014), we choose coefficients of the polynomial so that the stability domain expands along the imaginary axis and becomes singly connected. It provides better stability properties to rounding errors whereas the stability interval length reduces insignificantly.

For each k, $1 \le k \le m$, γ_k represents the length of such a maximal interval $[\gamma_k, 0]$, that for any $z \in [\gamma_k, 0]$ inequality $|Q_k(z)| \le 1$ satisfies. Taking into account, that $z = h\lambda$, in (13) for all $Q_k(z)$, $1 \le k \le m$ we replace h with $(h\gamma_k / \gamma_m)$. As a result, formula (13) may be written as follows

$$\begin{aligned} Q'_k(z) &= 1 + \sum_{i=1}^{k} c'_{ki} z^i, \\ c'_{ki} &= (\gamma_k / \gamma_m)^i c_{ki}, \quad 1 \le k \le m. \end{aligned} \tag{14}$$

The replacement of h with $(h\gamma_k / \gamma_m)$ means that the approximate solution obtained by intermediate schemes (5) is computed at points $(t_n + c'_{k1}h)$, $1 \le k \le m-1$, instead of $(t_n + c_{k1}h)$, $1 \le k \le m-1$. In this case the maximal stepsize, obtained according to the stability requirements of the basic scheme is also maximal for intermediate numerical formulas.

Determine coefficients of methods (5) as follows. First, using (Hairer and Wanner, 1996) we compute coefficients of polynomials (13), satisfying some defined properties. Further, compute coefficients of polynomials (14) applying corresponding substitution of variables. Taking into account, that elements of (k +

1)-th column of matrix B_m coincide with coefficients of the stability polynomials $Q'_k(z)$, form matrix

$$B_m = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 0 & c'_{11} & c'_{21} & \dots & c'_{m-1,1} \\ 0 & 0 & c'_{22} & \dots & c'_{m-1,2} \\ & & \dots & & \\ 0 & 0 & 0 & \dots & c'_{m-1,m-1} \end{pmatrix}, \quad (15)$$

Using in (11) vector $c'_k = (c'_{k1}, \dots, c'_{kk})^{\mathrm{T}}$ instead of c_k, we unambiguously determine all coefficients of methods (5) with conformed stability domains from linear system (9) and (11).

4 Accuracy and Stability Control

We use the estimation of local truncation error $\delta_{n,1}$ to control accuracy of the first order methods. Applying (12) we get that for the m-stage method it has the form

$$\delta_{n,1} = (0.5 - \sum_{j=2}^{m} b_{2j} p_{mj}) h^2 ff + O(h^3) = (0.5 - c_{m2}) h^2 ff + O(h^3),$$

where c_{m2} is the coefficient at z^2 in stability polynomial (8). Estimation $\varepsilon_{n,1}$ of the error can be computed using the formula

$$\varepsilon_{n,1} = [(0.5 - c_{m2}) / (\alpha_i - \alpha_j)](k_i - k_j), \quad 1 \le i, j \le m, \ i \ne j. \quad (16)$$

The graph of a solution of a stiff problem can be divided into two types of regions. The first one is the settling region (where values of solution derivatives are low), and the second one is the transition region (where values of solution derivatives are high). Taking this into account, to increase the performance of calculations we proceed as follows. We apply

$$\varepsilon'_{n,1} = [(0.5 - c_{m2}) / \alpha_2](k_2 - k_1). \quad (17)$$

to make an over-cautious estimation. As k_1 linearly depends on integration stepsize, omission of inequality of $\|\varepsilon'_{n,1}\| \le \varepsilon$ leads just to one additional computation of the right part of (3). Here, ε is the absolute or relative tolerance of calculations, $\|\cdot\|$ is some norm in R^N. Taking into account, that

$$hf(y_{n+1}) - k_1 = h^2 f'_n f_n + O(h^3),$$

the final decision on accuracy we make checking inequality $\|\varepsilon''_{n,1}\| \le \varepsilon$, where

$$\varepsilon''_{n,1} = (0.5 - c_{m2})(hf(y_{n+1}) - k_1). \quad (18)$$

We derive the inequality for stability control similarly to (Hairer and Wanner, 1996). To obtain this inequality we apply method (5) to problem (3) on $f(y) = Ay + b$, where A and b are N-dimensional matrix and vector with constant elements, respectively. As the result, we can estimate maximal eigenvalue λ_n^{max} of Jacobi matrix $\partial f(y_n)/\partial y$ of (3) using the formula

$$h\lambda_n^{\max} = |\alpha_2 \beta_{32}|^{-1} \max_{1 \le j \le N} \left| \frac{[\alpha_2 k_3 + \alpha_3 k_2 - (\alpha_2 + \alpha_3) k_1]_j}{[k_2 - k_1]_j} \right|. \quad (19)$$

Then, inequality for stability control for m-stage method (5) has form $h\lambda_n^{max} \le |\gamma_m|$, where $|\gamma_m|$ is stability interval length of the m-stage scheme.

5 First Order Method

For numerical solution of Cauchy problem (1) we consider the explicit five-stage Runge-Kutta method

$$\begin{aligned} y_{n+1} &= y_n + p_1 k_1 + p_2 k_2 + p_3 k_3 + p_4 k_4 + p_5 k_5, \\ k_1 &= hf(y_n), k_2 = hf(y_n + \beta_{21} k_1), \\ k_3 &= hf(y_n + \beta_{31} k_1 + \beta_{32} k_2), \\ k_4 &= hf(y_n + \beta_{41} k_1 + \beta_{42} k_2 + \beta_{43} k_3), \\ k_5 &= hf(y_n + \beta_{51} k_1 + \beta_{52} k_2 + \beta_{53} k_3 + \beta_{54} k_4), \end{aligned} \quad (20)$$

where y and f are real N-dimensional vector functions, t is an independent variable, h is the integration step, k_1, k_2, k_3, k_4, and k_5 are stages of the method, p_1, p_2, p_3, p_4, p_5, β_{21}, β_{31}, β_{32}, β_{41}, β_{42}, β_{43}, β_{51}, β_{52}, β_{53}, β_{54} are numerical coefficients, defining accuracy and stability properties of (20).

We choose coefficients of (20) so that it has the first accuracy order and the extended stability domain. The stability domain of a method with the maximal length of the stability interval is almost multiconnected. We design polynomials of the first, second, third, fourth, and fifth degree so that the corresponding them methods have singly connected stability domains with the stability interval close to the maximal possible one (see Figure 1).

Applying the algorithm from (Novikov and Rybkov, 2014), we get coefficients

$$c_{11} = c_{21} = c_{31} = c_{41} = c_{51} = 1,$$
$$c_{22} = 0.128025128205128,$$

$$c_{32} = 0.152092927269786, c_{33} = 0.00580524400854353,$$

$$c_{42} = 0.160464544241005, c_{43} = 0.00827164513740441,$$
$$c_{44} = 0.000133419220894335,$$

$$c_{52} = 0.164341322127141, c_{53} = 0.00948975952580473,$$
$$c_{54} = 0.000223956930863224, c_{55} = 1.85097275222353 \cdot 10^{-6}.$$

At that,

$$\gamma_1 = -2, \ \gamma_2 = -7.79, \ \gamma_3 = -17.46,$$
$$\gamma_4 = -30.99, \gamma_5 = -48.39.$$

Writing and resolving linear systems (9) and (11) using (15), we obtain the coefficients of method (20)

$$\beta_{21} = 0.0413243016210550, \beta_{31} = 0.0805823881610573,$$
$$\beta_{32} = 0.0805823881610573, \beta_{41} = 0.1191668151228434,$$
$$\beta_{42} = 0.1597820013984078, \beta_{43} = 0.0819394878966193,$$
$$\beta_{51} = 0.1570787892802991, \beta_{52} = 0.2379583021959820,$$
$$\beta_{53} = 0.1631711307360486, \beta_{54} = 0.0822916178203657,$$
$$p_1 = 0.1945277188657676, p_2 = 0.3151822878089125,$$
$$p_3 = 0.2437005934695969, p_4 = 0.1641555613805598,$$
$$p_5 = 0.0824338384751631.$$

To control accuracy of the numerical formula we use estimations (17) and (18). The stability interval length of numerical scheme (20) of the first accuracy order equals 17.46. Therefore, for its stability control we can apply inequality $h\lambda_n^{max} \le 17.46$, where $h\lambda_n^{max}$ is defined by formula (19).

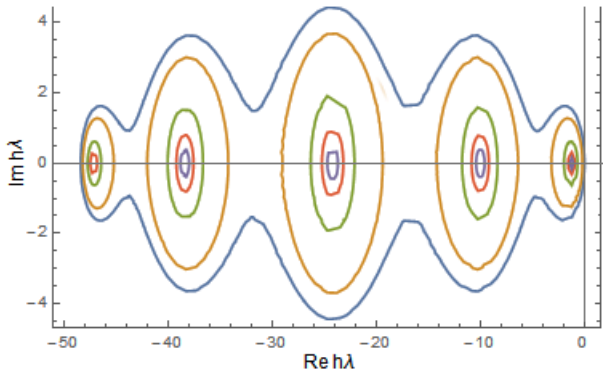


Figure 1. Stability domain of method (20).

6 Merson Method

The fourth accuracy order Merson method (Merson, 1957)

$$y_{n+1} = y_n + \frac{1}{6}k_1 + \frac{2}{3}k_4 + \frac{1}{6}k_5,$$
$$k_1 = hf(y_n), k_2 = hf(y_n + \frac{1}{3}k_1),$$
$$k_3 = hf(y_n + \frac{1}{6}k_1 + \frac{1}{6}k_2), \quad (21)$$
$$k_4 = hf(y_n + \frac{1}{8}k_1 + \frac{3}{8}k_3),$$
$$k_5 = hf(y_n + \frac{1}{2}k_1 - \frac{3}{2}k_3 + 2k_4),$$

is one of the most efficient and widely used explicit Runge-Kutta methods. The fifth computation of function f does not result in the fifth order of accuracy, but allows to extend the stability interval length to 3.5 and estimate truncation error $\delta_{n,4}$ using stages k_i, i.e.

$$\delta_{n,4} = (2k_1 - 9k_3 + 8k_4 - 2k_5)/30.$$

We apply inequality $\|\delta_{n,4}\| \le 5\varepsilon^{5/4}$ for accuracy control. The inequality is obtained assuming that the global error accumulated with local truncation errors (Novikov, 1997). Despite the fact that the inequality for accuracy control is obtained on a linear equation, it shows high reliability on solving non-linear problems.

Now let us derive the inequality for stability control. Applying to $k_3 - k_2$ the first order Taylor's formula with the remainder term written in the Lagrangian form, we have

$$k_3 - k_2 = h[\partial f(\mu_n)/\partial y](k_2 - k_1)/6,$$

where vector μ_n is computed in some vicinity of solution $y(t_n)$. Taking into account, that

$$k_2 - k_1 = h^2 f_n' f_n / 3 + O(h^3),$$

the inequality

$$v_{n,4} = 6 \cdot \max_{1 \le j \le N} \left| \frac{k_3^j - k_2^j}{k_2^j - k_1^j} \right| \le 3.5$$

can be used for stability control of (21), where 3.5 is the approximate length of stability interval (see Figure 2). Let $\varepsilon_{n,4} = \delta_{n,4}/5$. Then inequalities $\varepsilon_{n,4} \le 5\varepsilon^{5/4}$ and $v_{n,4} \le 3.5$ can be applied respectively for accuracy and stability control of scheme (21).

As estimation of eigenvalue $v_{n,4} = h\lambda_n^{max}$ is rough, stability control is used to limit integration stepsize and to switch between methods.

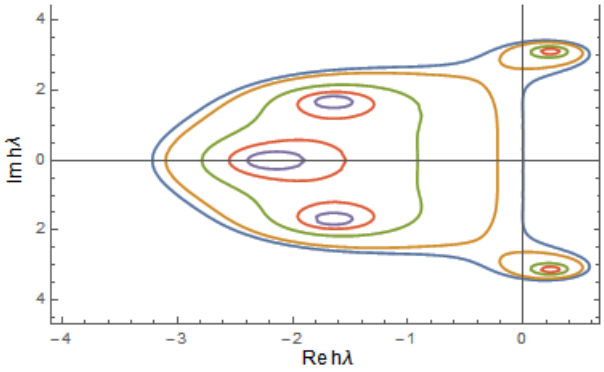


Figure 2. Stability domain of method (21).

The predicted step h_{n+1} is computed as follows. Step h^{ac}, that is chosen according to the requirements of accuracy, is computed using formula $h^{ac} = q_1 h_n$, where h_n is the latest accepted stepsize, and q_1, taking into account relation $\varepsilon_{n,4} = O(h_n^5)$, defined by $q_1^5 \varepsilon_{n,4} \le \varepsilon$. We compute step h^{st}, that is chosen according to the stability requirements, using $h^{st} = q_2 h_n$, where q_2, is defined by $q_2 v_{n,4} = 3.5$ as $v_{n,4} = O(h_n)$. Then, the predicted step h_{n+1} is computed using the formula

$$h_{n+1} = \max[h_n, \min(h^{ac}, h^{st})].$$

The given formula stabilizes stepsize over the settling region, where stability has the defining role.

7 Integration Algorithm

The algorithm of alternating order and step can be easily formulated on a base of the developed methods. It chooses the most efficient scheme on an each step. Calculations are always begun with the Merson method as it is more accurate. Switch to the first order method with conformed stability domains is performed on

omission of $v_{n,4} \le 3.5$. Transition to the Merson method is performed, if $v_{n,1} \le 3.5$ satisfies.

The norm in inequality for accuracy control is computed using the formula

$$\| \xi \| = \max_{1 \le i \le N} \frac{| \xi_i |}{| y_n^i | + r},$$

where i is a component number, r is a positive parameter. If inequality $\|y_n^i\| \le r$ satisfies for i-th component of a solution, absolute tolerance $r\varepsilon$ is controlled, otherwise, relative tolerance ε. On calculations r was assumed to be equal to 3.

8 Medical Akzo Nobel Problem

We chose the Medical Akzo Nobel problem (Mazzia and Magherini, 2008) to test our method. The Akzo Nobel research laboratories formulated this problem in their study of the penetration of radio-labeled antibodies into a tissue that has been infected by a tumor. This study was carried out for diagnostic as well as therapeutic purposes.

In (Mazzia and Magherini, 2008) there is considered a reaction diffusion system in one spatial dimension:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - kuv\,, \quad \frac{\partial v}{\partial t} = -kuv\,, \qquad (22)$$

which originates from the chemical reaction $A + B \rightarrow C$. Here, A, the radio-labeled antibody, reacts with substrate B, the tissue with the tumor, and k denotes the rate constant. The concentrations of A and B are denoted by u and v, respectively.

Making necessary transformations and defining $y(t)$ by $y = (u_1, v_1, u_2, v_2, \ldots, u_N, v_N)^T$ it is possible to write (22) in the form

$$\frac{dy}{dt} = f(t, y), \quad y(0) = g, \quad y \in R^{2N}, \quad 0 \le t \le 20\,, \qquad (23)$$

Here, the integer N is a user-supplied parameter. The function f is given by

$$f_{2j-1} = \alpha_j \frac{y_{2j+1} - y_{2j-3}}{2\Delta\zeta} + \beta_j \frac{y_{2j-3} - 2y_{2j-1} + y_{2j+1}}{(\Delta\zeta)^2} - k y_{2j-1} y_{2j},$$

$$f_{2j} = -k y_{2j} y_{2j-1},$$

where

$$\alpha_j = 2(j\Delta\zeta - 1)^3 / c^2\,,\ \beta_j = (j\Delta\zeta - 1)^4 / c^2\,,\ 1 \le j \le N\,,$$

$$\Delta\zeta = 1/N\,,\ y_{-1}(t) = \varphi(t)\,,\ y_{2,N+1} = y_{2,N-1}\,,\ g \in R^{2N}\,,$$

$$g = \left(0, v_0, 0, v_0, \ldots, 0, v_0\right)^T.$$

The function φ(t) = 2 at 0 < t ≤ 5 and φ(t) = 0 at 5 < t ≤ 20. Values for the parameters k, v0, and c are 100, 1, and 4, respectively. Graph of the time and space dependencies of u and v is shown in Figure 3.

9 Numerical Results

Calculations were performed on Intel(R) Core(TM) i3-5010U CPU with double precision. The parameter N was equal 200 that means that the system to be solved involved 400 equations.

The stiffness ratio of the Medical Akzo Nobel problem approximately equals 10^6. The graph of 133rd component of the solution is shown in Figure 4.

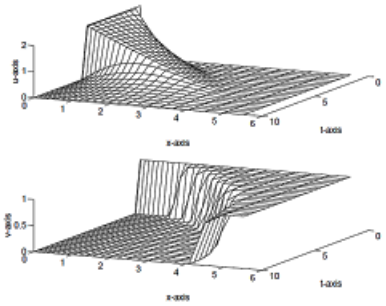


Figure 3. u and v as functions of time and space.

Below *IS*, *IW*, and *IF* represent, respectively, total numbers of steps, declined solutions (due to omission of the defined absolute tolerance), and computed right parts of the problem.

The algorithm of alternating order and step based on the first order method with conformed stability domains and the Merson method with accuracy and stability control gives the following results. For the defined absolute tolerance equal to 10^{-4} we have IS = 11 505, IW = 1 266, and IF = 70 893. For the absolute tolerance 10^{-7}: IS = 72 658, IW = 10 333, and IF = 403 066.

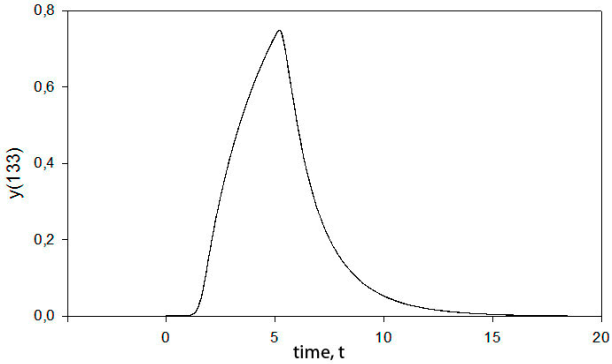


Figure 4. Solution of the Medical Akzo Nobel problem.

10 Conclusions

From the numerical results it follows that stability control leads to the efficiency gain due to the reduction of some declined solutions appearing as a result of instability of a numerical formula. Simulation of other test examples gives similar tendency. The designed method is aimed at the solution of large-scale problems of moderate stiffness with low accuracy, as well as problems with protensive settling regions, where the first order methods with conformed stability domains give growth of the efficiency.

The constructed algorithm is designed for low precision calculations – about 1% and lower. In this

case, its maximum efficiency is reached. In the algorithm, with its parameters, one can specify different modes of calculations: 1) with the explicit first order method with conformed stability domains either with or without stability control; 2) with the Merson method either with or without stability control; 3) with automatic choice of a numerical scheme. Therefore, this algorithm can be applied both for solving stiff and non-stiff problems. In calculations with automatic choice of a numerical scheme, the integration algorithm makes a decision whether a problem to be solved is stiff or not by itself.

Acknowledgements

This work is partially supported by Russian Foundation of Fundamental Researches (project №17-07-01513A).

References

G. Dahlquist: A Special Stability Problem for Linear Multistep Methods. *BIT*, 3: 23–43, 1963.

E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*. Springer-Verlag, 1996.

F. Mazzia and C. Magherini. *Test set for Initial Value Problem Solvers*. University of Bari and INdAM, Research Unit of Bari. Release 2.4, 2008.

E. A. Novikov. *Explicit Methods for Stiff Systems*. Novosibirsk: Nauka, 1997. (in Russian)

E. A. Novikov and M. V. Rybkov: Numerical Algorithm for Stability Domains Design for the First Order Methods. *Vestnik Buryatskogo Gosudarstvennogo Universiteta*, 9(2): 80-85, 2014. (in Russian)

E. A. Novikov and Y. V. Shornikov. *Computer Simulation of Hybrid Stiff Systems*. Novosibirsk: Publishing House of the Novosibirsk State Technical University, 2012. (in Russian)

R. H. Merson: *An Operational Method for Integration Processes*. Proc. Symp. Data Processing. Weapons Research Establishment, 331 p, 1957.

Numerical Algorithm for Design of Stability Polynomials for the First Order Methods

Eugeny A. Novikov[1] Mikhail V. Rybkov[2] Anton E. Novikov[3]

[1]Institute of Computational Modelling, Federal Research Center, Russia, novikov@icm.krasn.ru
[2]Institute of Mathematics and Fundamental Informatics, Siberian Federal University, mixailrybkov@yandex.ru
[3]Institute of Mathematics and Fundamental Informatics, Siberian Federal University, Russia, aenovikov@bk.ru

Abstract

This paper derives an algorithm for computing coefficients for stability polynomials of a degree up to $m = 35$. These coefficients correspond to explicit first order Runge-Kutta methods. Authors showed dependence between stability polynomial values at extreme points and both size and form of a stability domain. Numerical results are given.

Keywords: stiff problem, explicit methods, stability polynomials

1 Introduction

Heterogeneous algorithms are applied to solving stiff problems in a number of situations. Such algorithms are designed using the fact that on the settling and transition regions integration stepsizes are limited according to the requirements of stability and accuracy, respectively. Efficiency growth is achieved by applying an explicit scheme over the transition region and an L-stable scheme over the settling region. The switch between methods is performed using an inequality for stability control.

The problem is that the size of stability domains of the known methods is too small. Some monographs and papers present explicit methods with extended stability domains (Novikov, Shornikov, 2012). The way to obtain stability polynomials providing the maximal length of a stability domain is considered in (Skvortzov, 2011). Novikov (1997) proposed an algorithm which provides polynomials coefficients. The algorithm allows to design explicit Runge-Kutta methods with specified stability domain forms and sizes. Furthermore, coefficients of stability polynomials having a degree up to $m = 13$ are found there.

Here we develop an algorithm that provides stability polynomials coefficients having a degree up to $m = 27$. The coefficients correspond to explicit first order Runge-Kutta methods. It is shown that the form, size, and structure of a stability domain depend on the position of the stability polynomial roots on the complex plane.

2 Explicit Runge-Kutta Methods

To solve the stiff problem

$$y' = f(t,y),\ y(t_0) = y_0,\ t_0 \le t \le t_k,$$

where y and f are smooth real N-dimensional vector-functions, t is an independent variable, in (Novikov, 1997) explicit methods

$$y_{n+1} = y_n + \sum_{i=1}^{m} p_{m\,i} k_i,$$
$$k_i = hf\left(t_n + \alpha_i h, y_n + \sum_{j=1}^{i-1} \beta_{ij} k_j\right)$$

are considered, where k_i, $1 \le i \le m$, are stages of the method, h is the integration stepsize, p_{mi}, α_{ij}, and β_{ij} are numerical coefficients defining accuracy and stability properties of this numerical scheme. For simplicity, let us consider the following Cauchy problem for the autonomous system of ODEs

$$y' = f(y),\ y(t_0) = y_0,\ t_0 \le t \le t_k. \quad (1)$$

We apply methods of the form

$$y_{n,i} = y_n + \sum_{j=1}^{i} \beta_{i+1,j} k_j, 1 \le i \le m-1,$$
$$y_{n+1} = y_n + \sum_{i=1}^{m} p_{m\,i} k_i, \quad (2)$$

to solve (1), where $k_i = hf(y_{n,i-1})$, $1 \le i \le m$, $y_{n,0} = y_n$. All the findings those are to obtained below can be used for non-autonomous problems, if

$$\alpha_1 = 0,\ \alpha_i = \sum_{j=1}^{i-1} \beta_{ij},\ 2 \le i \le m.$$

Stability of one-step methods is widely studied on the Dahlquist equation $y' = \lambda y$, $y(0) = y_0$, $t \ge 0$ with complex λ, $\mathrm{Re}(\lambda) < 0$ (Dahlquist, 1963). Applying the second formula from (2) to solve $y' = \lambda y$, we get

$$y_{n+1} = Q_m(z) y_n,\ Q_m(z) = 1 + \sum_{i=1}^{m} c_{m\,i} z^i,$$
$$c_{mi} = \sum_{j=i}^{m} b_{ij} p_{m\,j},\ 1 \le i \le m,$$

where $z = h\lambda$. Hence, the stability function of a m-stage explicit Runge-Kutta method is polynomial $Q_m(z)$ of a degree m. Novikov (1997) gave order conditions for methods of form (2) and, in particular, method (2)

has the first accuracy order, if $p_{m1} + \ldots + p_{mm} = c_{m1} = 1$. Further, we consider the problem of finding such coefficients that a stability domain has specified form and size.

3 Stability Polynomials Over $[\gamma_m, 0]$

Let k and m be given integers, $k \le m$. Consider polynomials

$$Q_{m,k}(x) = 1 + \sum_{i=1}^{k} c_i x^i + \sum_{i=k+1}^{m} c_i x^i , \qquad (3)$$

where c_i, $1 \le i \le k$, are defined, and c_i, $k+1 \le i \le m$, are arbitrary. Usually c_i, $1 \le i \le k$, are determined according to the requirements of accuracy. Therefore, let us assume that $c_i = 1/i!$, $1 \le i \le k$.

Denote extreme points of (3) by $x_1, \ldots, x_{m-1}$, at that $x_1 > x_2 > \ldots > x_{m-1}$. Define unknown coefficients c_i, $k+1 \le i \le m$, so that polynomial (3) has predefined values at extreme points x_i, $k \le i \le m-1$, i.e. $Q_{m,k}(x_i) = F_i$, $k \le i \le m-1$, where $F(x)$ is some given function, $F_i = F(x_i)$. For this purpose, consider the following system of algebraic equations

$$\begin{gathered} Q_{m,k}(x_i) = F_i, Q'_{m.k}(x_i) = 0, k \le i \le m-1, \\ Q'_{m.k} = \sum_{i=1}^{m} i c_i x^{i-1}, \end{gathered} \qquad (4)$$

in variables x_i, $k \le i \le m-1$, and c_j, $k+1 \le j \le m$.

Rewrite (4) in the form, suitable for calculations on the computer. Denote through y, z, g, and r vectors with components

$$y_i = x_{k+i-1}, \; z_i = c_{k+i}, \; g_i = F_{k+i-1} - 1 - \sum_{j=1}^{k} c_j y_i^j ,$$

$$r_i = -\sum_{j=1}^{k} j c_j y_i^{j-1} , \; 1 \le i \le m-k ,$$

through E_1, E_2, E_3 – diagonal matrices with elements

$$e_1^{ii} = k + i , \; e_2^{ii} = 1 / y_i ,$$

$$e_3^{ii} = (-1)^{k+i-1}, \quad 1 \le i \le m-k,$$

and through A – a matrix with elements $a_{ij} = y_i^{k+j}$, $1 \le i, j \le m-k$. Using these notations problem (4) can be written as follows

$$Az - g = 0, \; E_2 A E_1 z - r = 0 . \qquad (5)$$

System (5) is ill-conditioned that leads to some difficulties while solving it with the fixed point iteration method. For convergence of the Newton's method, it is necessary to somehow obtain good initial values that in this case is a separate difficult problem. If we assume in (4) that $F_i = (-1)^i$, $k \le i \le m-1$, we find the polynomial with the maximal length of the stability interval. In this case the problem of computation of initial value y^0 is solved using values of the Chebyshev polynomial at extreme points over interval $[-2m^2, 0]$, where m is a degree of polynomial (3). Those values can be computed using the formula

$$y_i = m^2 [\cos(i\pi / m) - 1] , \; 1 \le i \le m-1 . \qquad (6)$$

Substituting (6) in the system (5), get coefficients of the Chebyshev polynomial, for those $|Q_{m1}(x)| \le 1$ on $x \in [-2m^2, 0]$. For any k we can take (6) as initial values and, as numerical results show, there is good convergence rate in this case. If $F_i \ne (-1)^i$, $k \le i \le m-1$, then the choice of initial values is a separate difficult problem.

Let us describe a way to solve (5) that does not require good initial values. Apply the relaxations for the numerical solution of (5). The main idea of the relaxations is that for a steady-state problem we run unsteady-state process which solution settles to the solution of the initial problem. Consider the Cauchy problem

$$y' = E_3 (E_2 A E_1 A^{-1} g - r), \; y(0) = y_0 . \qquad (7)$$

Apparently, after a stationary point of (7) has been found, the stability polynomial coefficients can be computed from the system (5). Notice, that due to using matrix E_3 all eigenvalues of the Jacobi matrix of (7) have negative real components, i.e. problem (7) is stable. Numerical results show that (7) is a stiff problem. Applying methods that require evaluation of the Jacobi matrix may cause difficulties while solving (7). Therefore, we solve (7) with the second accuracy order method that uses numerical computing and freezing the Jacobi matrix (Novikov, 2008). When applied to the problem $y' = f(y)$, $y(0) = y_0$, this method takes the form

$$\begin{gathered} y_{n+1} = y_n + a k_1 + (1-a) k_2 , \; D_n = E - a h_n A_n , \\ D_n k_1 = h_n f(y_n), D_n k_2 = k_1 . \end{gathered} \qquad (8)$$

Here, $a = 1 - 0.5\sqrt{2}$, k_1 and k_2 are stages of the method, E is the identity matrix, h_n is the integration stepsize, A_n is a matrix representable in the form $A_n = f_n' + h_n R_n + O(h_n^2)$, $f_n' = \partial f(y_n)/\partial y$ is the Jacobi matrix of (7), R_n is the integration stepsize independent matrix. Since matrix R_n is arbitrary, problems of numerical solving and freezing the Jacobi matrix can be concerned simultaneously. To control accuracy of (8) we apply the inequality

$$\varepsilon(j_n) = \| D_n^{1-j_n} (k_2 - k_1) \| \le a\varepsilon / |a - 1/3| , \; 1 \le j_n \le 2 , \qquad (9)$$

where ε is the required accuracy of calculations, $\|\cdot\|$ is some norm in R^N, and integer variable j_n is chosen minimum for which inequality (9) is satisfied. The numerical differentiation step s_j, $1 \le j \le N$, is chosen using the formula $s_j = \max\{10^{-14}, 10^{-7}|y_j|\}$. In this case j-th column a_n^j of matrix A_n is computed using the formula

$$a_n^j = [f(y_1, \ldots, y_j + s_j, \ldots, y_N) - f(y_1, \ldots, y_j, \ldots, y_N)] / s_j ,$$

$$1 \le j \le N,$$

i.e. it is required to perform N computations of the right part of problem (7) to define A_n. An attempt to use previous matrix D_n is performed after each successful integration step. To preserve stability properties of the numerical scheme, on freezing matrix D_n the integration stepsize is kept permanent. Recomputation of the matrix is carried out in the following cases: 1) calculations accuracy is

degenerated, 2) quantity of steps with a frozen matrix has reached maximal number I_h, 3) the predicted step is greater than the previous successful one in Q_h times.

4 Stability Polynomials Over [–1, 1]

It is not difficult to see that stability polynomial coefficients approach zero as m increases. Novikov (1997) presented coefficients c_i, $k + 1 \le i \le m$, for polynomials of a degree up to $m = 13$. Now consider an algorithm providing polynomials with specified properties over the interval [−1, 1]. In this case coefficients c_i grow not that much, and it is possible to derive polynomials for $m > 13$. Denote through $|\gamma_m|$ the length of stability interval of m-stage explicit formula of the Runge-Kutta type, i.e. the inequality $|Q_{m,k}(x)| \le 1$ over the interval $[\gamma_m, 0]$ is satisfied. Then, substituting $x = 1 - 2z/\gamma_m$ we can map $[\gamma_m, 0]$ into [−1, 1] and obtain polynomial

$$Q_m(z) = \sum_{i=0}^{m} d_i z^i \ . \qquad (10)$$

Coefficients d_i, $0 \le i \le m$ of polynomial (10) and coefficients c_i, $0 \le i \le m$, of (3) satisfy the relation

$$c = UVd, \qquad (11)$$

where $d = (d_0, \ \dots \ , d_m)^T$, $c = (c_0, \ \dots \ , c_m)^T$, U is a diagonal matrix with elements $u^{ii} = (-2/\gamma_m)^{i-1}$, $1 \le i \le m + 1$. Elements v^{ij} of V are defined by

$$v^{1j} = 1 \ , \ 1 \le j \le m+1 \ ; \ v^{ij} = v^{i,j-1} + v^{i-1,j-1} \ ,$$

$$2 \le i \le j \le m+1 \ ; \ v^{ij} = 0 \ , \ i > j \ .$$

Obviously, V represents the Pascal's triangle which elements are easily computed using a recurrent formula. Therefore, after deriving the polynomial (10) over interval [–1, 1], using (11) it is easy to compute coefficients of the polynomial (3).

Now let us derive polynomial (10). We denote the extreme points of (10) through $z_1, \ \dots \ , z_{m-1}$, at that $z_1 > z_2 > \ \dots \ > z_{m-1}$. We compute coefficients d_i, $0 \le i \le m$, under condition that polynomial (10) has predefined values in extreme points z_i, $1 \le i \le m - 1$, i.e.

$$Q_m(z_i) = F_i, \ 1 \le i \le m-1 \ ,$$

where $F(z)$ is some given function, $F_i = F(z_i)$. For that, consider the following system of algebraic equations

$$Q_m(z_i) = F_i, Q'_m(z_i) = 0, 1 \le i \le m-1 \ , \qquad (12)$$
$$Q'_m(z) = \sum_{i=1}^{m} i d_i z^{i-1},$$

here the normality conditions $Q_m(-1) = (-1)^m$ and $Q_m(1) = 1$ are satisfied.

Rewrite (12) in the form, suitable for calculations on the computer. For this purpose, denote by y, w, g, and r vectors with components

$$y_j = z_j \ , \ r_j = 0 \ , \ 1 \le j \le m-1 \ ; \ w_i = d_{i-1} \ , \ 1 \le i \le m+1 \ ,$$

$$g_i = F_i \ , \ 1 \le i \le m-1 \ ; \ g_i = 1 \ , \ i = m \ ;$$

$$g_i = (-1)^m \ , i = m+1 \ ;$$

through E_1 and E_2 matrices of dimension $(m + 1)\times(m + 1)$ and $(m - 1)\times(m + 1)$, respectively, with elements of the form

$$e_1^{ij} = j - 1 \ , \ 1 \le j \le m+1 \ ; \ e_2^{ii} = 1 / y_i \ , \ 1 \le i \le m-1 \ ,$$

and through A – a matrix of dimension $(m + 1)\times(m + 1)$ with elements

$$a^{ij} = y_i^{j-1} \ , \ 1 \le i \le m-1 \ , \ 1 \le j \le m+1 \ ; \ a^{m,j} = 1 \ ,$$

$$a^{m+1,j} = (-1)^{j+1} \ , \ 1 \le j \le m+1 \ .$$

Now problem (12) can be written as follows

$$Aw - g = 0 \ , \ E_2 A E_1 w - r = 0 \ . \qquad (13)$$

For the numerical solution of (13) we use the relaxations (Novikov, 1997). After the determination of polynomial (10) coefficients, compute the coefficients of polynomial (3) using relation (11). Find value γ_m under assumption that the polynomial to be obtained corresponds to the first order method, i.e. $c_1 = 1$. Having written the second relation and having made necessary transformations, we get

$$\gamma_m = \left\{ -2 \sum_{j=1}^{m+1} v_{2j} d_j \right\} / c_1 \ , \ \gamma_m^0 = -2m^2 \ .$$

5 Form and Size of Stability Domains

Let us describe how the choice of values F_i affects the size and form of a stability domain. If we let $F_i = (-1)^i$, $k \le i \le m - 1$, then the stability interval length is known and computed using the formula $|\gamma_m| = 2m^2$. In this case for given m we get the maximal length of a stability domain along the real axis. Figure 1 shows level curves $|Q_{m,k}(x)| = 1$, $|Q_{m,k}(x)| = 0.8$, $|Q_{m,k}(x)| = 0.6$, $|Q_{m,k}(x)| = 0.4$, and $|Q_{m,k}(x)| = 0.2$ on the complex plane $\{h\lambda\}$ for the stability domain on $m = 4$, $k = 1$, $F = \{-1, 1, -1\}$. The stability interval length $|\gamma_m|$ of the corresponding method equals 32.

In case if the stability interval length is maximal, a stability domain is almost multiconnected, so rounding errors may lead to stepping out of the stability domain. To solve this problem we need to stretch the stability domain along the imaginary axis in tangency points of parts of the stability domain. For this purpose, we can let $F_i = (-1)^i \mu$, $1 \le i \le m - 1$, $0 < \mu < 1$. Numerical results show that if $\mu = 0.9$, the stability interval length becomes shorter by 5–8% comparing to the maximal possible length equal to $2m^2$. At that, the stability domain stretches along imaginary axis at the tangency points. This provides better stability properties of the corresponding method to rounding errors on insignificant reduction of the stability interval length. If we assume $\mu = 0.95$, then the stability interval length is reduced by 3–4%. The stability domain of the five-stage method on $\mu = 0.9$ is shown in Figure 2. The stability interval length of this method $|\gamma_m| = 30.00$.

As μ decreases from 1 to 0, roots of polynomial (3) get closer to each other on the real axis. Therefore, the stability interval length is reduced. The ellipsises,

which are well-defined on $\mu = 1$ get closer not providing essential stretch of the stability domain along the imaginary axis. Therefore, depending on the problem to be solved it is reasonable to choose value μ from 0.8 to 0.95.

On solving problems, which Jacobi matrices have eigenvalues with imaginary components and which solutions have an oscillating behavior, the extension of a stability interval is not always obligatory. In this case, the integration stepsize is rather small due to the accuracy requirements and thus it is more reasonable to extend a stability domain along the imaginary axis. If the Jacobi matrix have pure imaginary eigenvalues, it is necessary to have the condition $|Q_{m,k}(x)| = 1$ satisfied over some region on the imaginary axis. This requirement is satisfied as k increases.

For the first order methods, i.e. for $k = 1$, it is possible to make the requirement satisfied choosing appropriate values of function F. For instance, on $m = 4$, $k = 1$, $F = \{0.75, 0.80, 0.75\}$ we obtain a polynomial, satisfying this requirement (Figure 3). Since m is even and all the values F_i are positive, the graph of the polynomial does not cross the real axis, at that, polynomial has two pairs of complex conjugate roots. Therefore, the stability domain stretches along the imaginary axis and some region of the imaginary axis belongs to the stability domain. The stability interval length equals 2.89. On reducing values F_i the length of a stability domain along the real axis gets greater. While further reducement of values F_i the stability interval length $|\gamma_m|$ also becomes greater but the region on the imaginary axis belonging to the stability domain becomes less. Therefore, for developing first order methods aimed at solving oscillating problems, it is reasonable to choose stability polynomials which have a couple of complex conjugate roots in a complex plane $\{h\lambda\}$ nearby the origin of coordinates. At that, values F_i that correspond to these roots are needed to be chosen close to 1, so that the stability domain has the maximal region of the imaginary axis in it.

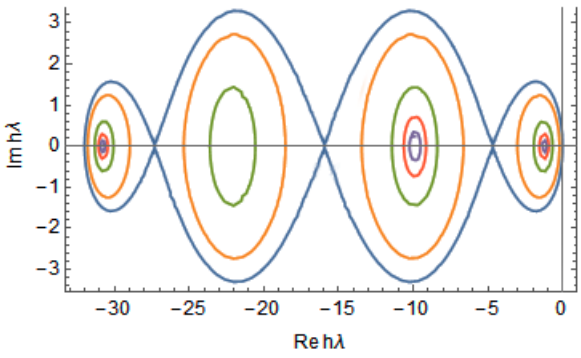


Figure 1. Stability domain on parameters m = 4, k = 1, F = {–1, 1, –1}.

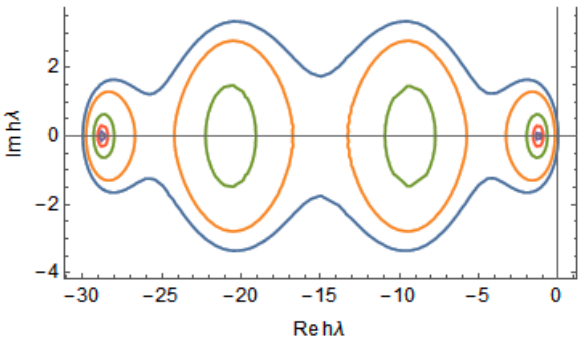


Figure 2. Stability domain on parameters m = 4, k = 1, F = {–0.9, 0.9, –0.9}.

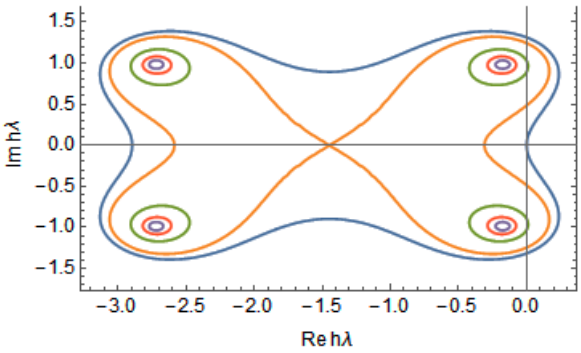


Figure 3. Stability domain on parameters m = 4, k = 1, F={0.75, 0.80, 0.75}.

6 Numerical Results

The numerical results show that coefficient c_m of polynomial (3) reduces as m grows. In particular, on $m = 13$ and $k = 1$ value c_m is of the order of 10^{-26}. It is difficult to solve problem (7) for $m > 13$ due to rounding errors. Numerical results of solving (11) show that polynomial (10) coefficients d_i, $0 \le i \le m$, grow in magnitude simultaneously with m. In particular, on $m = 13$ value $\max_{0 \le i \le m} |d_i|$ is of the order of 10^5, and on $m = 25$ of the order of 10^9, i.e. d_i grow slower than c_i. We transit from polynomial (10) coefficients to coefficients of (3) using (11) after (13) has been solved. This allows to compute the coefficients of stability polynomials of a degree up to $m = 27$.

It is difficult to solve problem (11) with double precision for $m > 27$ due to the appearing rounding errors. The algorithm using the "Quade-Double Precision Library" (described in (Hida, 2000)) was developed to compute the stability polynomials coefficients with greater m.

The "QD Precision Library" allows performing calculations with higher accuracy. While the standard data type 'double', allowing to represent numbers with double precision, is confined to 53 bits of the binary mantissa and provides precision about 16 decimal numerals, numbers of the data type 'dd_real' from the

library QD has the 106-bit mantissa that provides precision about 32 decimal numerals. In fact, the number of the type dd_real is the software-implemented sum of two numbers of the type 'double'. At that, the mantissa of the sum elongates in two times, but the range of values, presentable in new data type does not change and the possible values vary from about 10^{-308} to 10^{308}, as for the standard 'double'. Despite the confinement, accuracy of the representation of numbers in this diapason increases.

On the implementation of the algorithm for computation of the coefficients of (8) using the data type 'dd_real' the main input parameters of the algorithm − accuracy of calculations ε and differentiation stepsize s_j did not change. The Chebyshev polynomial values at the extreme points were chosen for initial conditions. The improved precision of the numbers representation allowed to compute polynomial coefficients for $m > 27$.

7 Conclusions

Authors of this article computed the coefficients for stability polynomials of a degree up to $m = 35$ using an algorithm providing polynomials over the interval $[-1, 1]$. These coefficients correspond to the first order Runge-Kutta methods with specified both form and size of stability domains. It is shown that the choice of values of stability polynomials at extreme points affects form and size of a stability domain. The proposed algorithm for designing stability domains increases the efficiency of explicit methods. Furthermore, it allows to develop algorithms of alternating order and step for solving problems of moderate stiffness. If the solution behavior of a problem which is to be solved is known, then it is possible to design an integration algorithm with the stability domain suitable for the given class of problems.

From our point of view, one of the main future applications of these results is to use the proposed algorithm to design numerical methods for solving ODEs systems. These methods can be included in libraries for software aimed at computer simulation.

Acknowledgements

This work is partially supported by Russian Foundation of Fundamental Researches (project №17-07-01513A).

References

G. Dahlquist: A Special Stability Problem for Linear Multistep Methods. *BIT*, 3: 23–43, 1963.

Y. Hida, X. S. Li, and D. H. Bailey. *Quad-Double Arithmetic: Algorithms, Implementation, and Application.* Technical Report LBNL-46996, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, 2000.

E. A. Novikov. *Explicit Methods for Stiff Systems*. Novosibirsk: Nauka, 1997. (in Russian)

E. A. Novikov and Y. V. Shornikov. *Computer Simulation of Hybrid Stiff Systems*. Novosibirsk: Publishing House of the Novosibirsk State Technical University, 2012. (in Russian)

A. E. Novikov and E. A. Novikov: L-stable (2,1)-method for solving stiff non-autonomous problems. *Computational Technologies*, 13: 477– 482, 2008. (in Russian)

L. M. Skvortzov: Simple way to Design Stability Polynomials for Explicit Stabilized Runge-Kutta Methods. *Matematicheskoe modelirovanie*, 23(1): 81–86, 2011.

Modelling and Simulation of PtG Plant Start-Ups and Shutdowns

Teemu Sihvonen[1] Jouni Savolainen[2] Matti Tähtinen[1]

[1]Renewable energy processes, VTT Technical Research Centre of Finland Ltd., Jyväskylä, Finland, {teemu.sihvonen, matti.tahtinen}@vtt.fi
[2]VTT Technical Research Centre of Finland Ltd., Espoo, Finland, jouni.savolainen@vtt.fi

Abstract

As the share of renewable energy sources increases the need for energy storages increases also due to fluctuating nature of renewables (solar and wind). Power-to-Gas (PtG) is one a promising energy storing concept. In PtG process renewable electric energy is used to produce hydrogen via electrolysis. Hydrogen is used together with carbon dioxide in methanation process to produce storable methane. Automation and operation logics of PtG plant have been in the minority in the literature. In this work we have studied start-up and shutdown logics of a PtG plant with dynamic simulations. With this approach we have identified development needs for such logics

Keywords: power-to-gas, automation, start-up, shutdown

1 Introduction

In the Paris Agreement 2015 governments have agreed to a long-term goal in global average temperature increase, global emission to peak as soon as possible and rapid emission reductions (UN, 2015). European commission has already 2011 proposed a target for renewable energy in EU's overall energy mix to be 20% by 2020 (EREC, 2011). This increases the need for wind and solar energy which are inconsistent energy forms which have negative effect on electric grid stability (Götz et al., 2016). This opens a need for long term large capacity electric energy storage.

One option for energy storing and electric grid stabilizing is the power-to-gas (PtG). The PtG works as a link between the power grid and the natural gas grid by converting electric energy to gas suitable to gas grid. The PtG is a two-step process, where firstly H_2 is produced by water electrolyser and secondly H_2 together with CO_2 is converted to CH_4 by a methanation reaction (dena German Energy Agency, 2015).

Even though the PtG process is usually driven as much as possible due to high investment costs, sometimes process might need to be shutdown. In the literature shutdowns and start-ups have been presented only for individual parts of the PtG process e.g. electrolysis (Eichman et al., 2014). Literature seems to lack of publications and research related to full dynamic simulations of PtG process from electric grid to gas grid. Experimental results for this process have been presented for example by (Zuberbüler et al., 2016). In this work, all parts of the process are modelled individually and the control logics for start-ups and shutdowns are linked to work together as a plant-level coordinating system. The main aim of this paper is to demonstrate the potential of using a dynamic process simulator in the development of such control logics for a PtG process. To the authors' best knowledge, the presented detailed simulation results are the first of a kind in the literature.

2 Methodology

Since start-ups and shutdowns are inherently dynamic operations, a dynamic simulation platform, Apros® (Savolainen et al., 2016; Silvennoinen et al., 1989; Tähtinen et al., 2016; Weiss et al., 2016) was used. The schematic flowsheet of the modelled PtG plant is shown in Figure 1. The PtG plant modelled in this work has three 3 MW electrolysers connected to power grid. Their electrical power consumption follows the grid frequency. Hydrogen and oxygen produced by electrolysis are flowing through interim storages before usage for buffering and compression reasons. Hydrogen is used in methanation while in this study oxygen does not have any specific usage. CO_2 for methanation comes from a constant flow power plant flue gases by MEA absorption. The captured CO_2 has also an interim storage mainly for compression reasons. The produced CH_4 from methanation reactor is compressed to natural gas grid pressures in a compression station.

All individual process components are described in upcoming subsections. More detailed description of the process model can be found from the study by

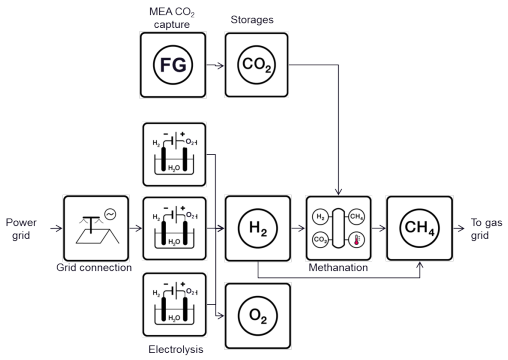


Figure 1. Flowsheet of the modelled PtG process.

Savolainen et al. (2016).

2.1 Power grid connection

The power grid connection block takes grid frequency in hertz as an input. Based on grid frequency the connection block gives an output electrical power for the electrolysers. This power is determined by user given droop curve in Figure 2. Maximum output power for one electrolyser was set to $3MW_e$.

2.2 Electrolyser

Three alkaline electrolysers working at 1 bar pressure and 70°C nominal temperature were used in this study. Electrolyser modelling was done with current-voltage (IV) curve from (Zhou and Francois, 2009). IV-curve is used to calculate the output molar flows of hydrogen and oxygen. Also released heat from the electrolyser cells is calculated and the needed cooling circulations with heat exchangers and major liquid and metal thermal masses are modelled. The alkaline electrolyser model was validated with industrial scale data and the relative errors were found to be within ±5% (Savolainen et al., 2016).

2.3 Interim gas storages

Interim gas storages for H_2, CO_2 and O_2 have similar construction in them, only pressure levels and the numbers of compression stages are different. In this study, the most important interim storage is the one for hydrogen. Interim H_2 storage has three different pressure levels connected to each other with semi-isothermal compression trains. More accurate description of interim H_2 storage can be found from (Tähtinen et al., 2016). The purpose of the interim storage is to stabilize gas flows from fluctuating sources to processes that need constant inlet gas flow, to provide storage for possible excess gas and to fill the gap when gas production is low. For example, hydrogen flow from electrolysis can be fluctuating heavily when electrolysis is used to stabilize power grid frequency or when intermittent energy sources are used.

2.4 Methanation

Hydrogen and carbon dioxide are mixed before entering to the methanation reactor. CO_2 feed from interim storage is controlled to achieve stoichiometric molar ratio (4:1 H_2:CO_2) for the methanation feed flow. A constant mixed feed flow from storages used is approximately 0.20 kg/s. Fixed bed methanation reactor has average pressure of 6.3 bar and the maximum temperature is controlled to be 550°C by steam cooling. A part from the output flow is recirculated back to the methanation reactor inlet. The amount of recirculation flow is controlled to by the quality of the gas produced in methanation. Aim for the CH_4 concentration on dry basis is set to 96 mol-%. The recirculated flow is cooled and the steam is condensed off before recirculation compressor. This flow is mixed with the gas mixture from the interim gas storages. The methanation reactor model was validated with laboratory measurements and was found to be in good agreement with them as indicated in Figure 3 below.

2.5 MEA CO_2 capture

Carbon dioxide used in methanation is captured from flue gas flow with a monoethanolamine (MEA) scrubber whose model is based on (Rao and Rubin, 2002). The MEA scrubber model calculates energy needs (reboiler duty, blower and pump electric energy) for CO_2 capture. Assumption in this model is that CO_2 source has excess production and is not limiting factor of the process model. This model has no actual dynamics but it can be used in the future to determine possible heat integrations of the process.

2.6 CH_4 compression

Product gas from methanation is cooled and steam is condensed off. After water removal the methane concentration is high enough for the natural gas grid (>95 mol-%). Methane is compressed to the natural gas grid pressure of 54 bar with a compression train consisting two compressor stages and water driven heat exchangers after them.

3 Control sequences

In this section, control logics for start-up and shutdown sequences are described. Specific control logics for each process component are presented.

3.1 CH_4 compression

Product gas from methanation is cooled and steam is condensed off. After water removal the methane concentra-

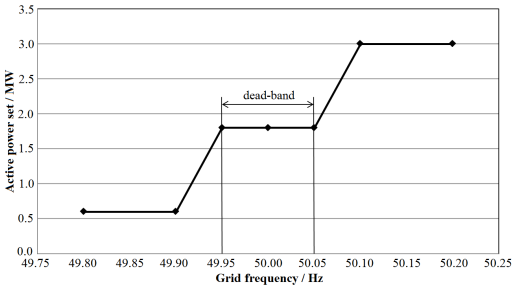


Figure 2. Droop curve of power grid connection.

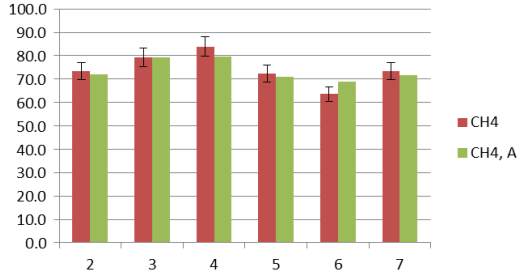


Figure 3. Comparison of product gas CH_4 mol-% between laboratory measured (CH_4) and simulated (CH_4, A).

tion is high enough for the natural gas grid (>95 mol-%). Methane is compressed to the natural gas grid pressure of 54 bar with a compression train consisting two compressor stages and water driven heat exchangers after them.

3.2 Start-up

The start-up is initiated from a cold state. Cold state start-up is the base case for start-up and it includes every step that is happening in a start-up. The stat-up order of equipment and the equipment specific routines as well as their inter-links are described in the following list.

3.2.1 Power grid connection

Since the time constants in power electronics are much smaller than in other process components there are not specific dynamic steps taken.

3.2.2 Methanation reactor

In its standby state, the methanation reactor is filled with hydrogen to prevent unwanted reactions between nickel catalyst and CO_2 (Jürgensen et al., 2015). During the start-up methanation reactor is firstly heated up to at least 275°C temperature before H_2 and CO_2 mixture can be fed in. This heating is done with an electric heater. During the start up the product quality does not meet the conditions for natural gas grid and recirculation is used to increase the product quality. Recirculation compressor starts when H_2 -CO_2 mixture starts to feed in the reactor. Even though the methane quality is low during the start-up some of the methanation product is fed to the natural gas grid. This is due to the fact that the hydrogen concentration up to 5vol-% is acceptable in the natural gas grid (Melaina et al., 2013). This also prevents pressure to rise in the methanation reactor uncontrollably high. Steam cooling of the reactor starts when methanation reactions cause temperature at the reactor inlet to rise close to 550°C.

3.2.3 Electrolysers

The three alkaline electrolysis units are identical and their start-up routine is the same. When alkaline electrolyser is shut off it is filled with nitrogen gas to prevent catalyst oxidation (Øystein et al., 2010). During start-up incoming water pushes nitrogen out from the cells. Water feed is opened when power feed from grid connection starts. Electric current cause's water to split into hydrogen and oxygen and the hydrogen prevents catalyst oxidation now.

Alkaline electrolysis start-up takes from 30 to 60 minutes (Øystein et al., 2010). In literature it is not specified how long it takes for removing the purge gas and how long it takes to heat up the electrolyser. For this reason in our model production restrictive part is taken to be the heating up of the equipment and nitrogen removal is not modelled. Production of H_2 and O_2 gases in the model starts instantaneously when power feed is turned on. This needs further validation as it is still unclear how production behaves during start-up. Part of the hydrogen should be used to prevent catalyst oxidation. The power feed to the electrolysers is started when temperature of methanation reactor is 230°C.

3.2.4 Interim gas storages

Interim gas storages have a control to keep the pressure at the inlet at the constant level. This control becomes active whenever there is an inlet flow causing the pressure rise. Outlet of the storages opens when downstream part of the process is ready for this. In other words, the interim hydrogen storage outlet opens when the methanation is heated up to the "ignition" temperature (at least 275°C) of methanation reaction. Next, CO_2 feed opens to fulfil stoichiometric molar feed ratio and finally CO_2 feed from MEA to interim storage is opened when storage outlet has been opened.

3.2.5 CH_4 compression

The compression train in CH_4 compression unit is turned on when methanation reactor is heated up. Feed to the natural gas grid can be started at this point even though CH_4 concentration is low and H_2 concentration is high. CH_4 compression is set to keep the pressure in the methanation reactor at 6.3 bars.

3.3 Shutdown

Shutdown sequence is a reverse process to start-up with some changes in actions of some process components.

3.3.1 Methanation reactor

Methanation reactor shutdown starts by extinguishing methanation reactions in the reactor. This is done by stopping CO_2 feed. Only pure hydrogen is feed to the methanation reactor until its H_2 concentration is close to 95 m-percent. Recirculation is closed and high H_2 concentration gas mixture is fed to the natural gas grid. When there is no CO_2 in the reactor it is safe to turn the cooling off. Additional cooling is not done for the reactor to reach the ambient temperature; this cooling is done only by natural convection.

3.3.2 CH_4 compression

Shutdown of methanation happens by decreasing the CO_2 concentration in the reactor. This causes decrease in the methnation reactions. The concentration of CH_4 decreases and H_2 increases. This is thought to not be a problem as it is allowed to feed some H_2 to the natural gas grid. CH_4 compressors first follow the inlet pressure and when H_2 fed to the methanation reactor is closed are CH_4 compressors also closed.

3.3.3 Interim gas storages

CO_2 feed to methanation is first to close during shutdown as explained in previous section. MEA capture is kept running until pressure vessels in CO_2 storage are full. H_2 feed to methanation is open until desired H_2 concentration in methanation reactor is reached. Compressors in H_2 and O_2 storages close as the feed from electrolysers stops.

3.3.4 Electrolysers

The model of alkaline electrolyser doesn't contain any real volume to where purge nitrogen could be fed during shutdown. For this reason the electrolyser shutdown is done by shutting the power feed from the power grid connection. The power feed is turned off when H_2 interim storage is full. H_2 storage is filled before total shutdown so that start-up of electrolyser and methanation can be done simultaneously. Cooling water feed is working as long as heat is produced in the electrolyser. Cooling to ambient temperature is left to be done by natural convection. In future development should be done to get the purge nitrogen feed for the model. Nitrogen gas flow works also as cooling agent which is beneficial for the electrolyser as oxidation is less severe in lower temperatures.

4 Results

In the experiment the start-up and shutdown sequences were activated by the operator, not process conditions. The start-up sequence was initiated after 10 minutes of steady down period. Star-up sequence was driven for 5 hours and 50 minutes. This time was long enough to stabilize the methanation process as can be seen from Figure 4.

The shutdown was started after the PtG plant had stabilized to normal mode, at the 6 hour mark in Figure 4-6. As can be seen, the shutdown process is much faster as electrolysers and methanation reactor were left to cool down only by natural convection. In total, 4 hours of shutdown was simulated to see at least some cooling in electrolysers, see Figure 5. Cooling of the methanation reactor was minimal due to the high total mass of the equipment and the insulation effect of steam cooling jacket.

Cold start-up of the methanation reactor is the slowest part of the PtG plant start-up. From the literature we did not find any time values for the methanation reactor's cold start-up. Heating ramp used in the simulations was selected so that too high temperature gradients would not happen, since a high temperature gradient could have negative effect on the catalyst durability. In the simulations the highest temperature gradients were found to happen when H_2 and CO_2 mixture feed was opened and closed, at 3 and 6 hour marks respectively in Figure 3. During start-up average temperature of the methanation reactor experienced 100°C drop in 3 minutes. Reason for this is the start-up of the recirculation and CH_4 compression units. Inlet gas is not heated before coming to the reactor and sudden rise on the flow cools the reactor. Also, rapid changes in temperature can result from changing reaction mixture composition. In other words, if the feed H_2:CO_2 molar ratio deviates from the stoichiometric value of 4:1, the reaction speeds and thus released heat is affected.

Fast changes in the gas feed causes also pressure fluctuations, which are detrimental for the durability of catalyst, in the methanation reactor, are seen at the lower part of Figure 4. During the start-up the methanation reactor experienced a pressure drop of 3.4 bars in a bit over 1.5 minutes.

During the start-up and steady operation of electrolysers the power from the power grid connection was fluctuating because grid frequency control was participated in. This did not cause high temperature gradients in the electrolyser and was found to be possible operation mode. Total heating up time for the electrolysers was 1 hour and 43 minutes, see Figure 5, top. Reason for long heating up time is the 1.8 MW average electric power used during the heating up period. Start-up period was also used for power grid frequency controlling. Start-up times from 30 to 60 minutes mentioned in literature could be achievable with maximum electric power of 3 MW.

The alkaline electrolysis H_2 and O_2 productions start instantaneously when electric power is switched on and the effect of temperature has low impact to this, Figure 5 lower part. This part of the electrolyser model needs further investigation and development.

Flow from the methanation to the natural gas grid has high changes in quality, measured as CH_4 molar-%, during the start-up and shutdown. During start-up this flow changes from a high H_2 concentration to high CH_4 concentration (and during shutdown vice versa), as can be seen in Figure 6. As mentioned in previous sections this is thought to be acceptable because the natural gas grid can take up to 5 vol-% of hydrogen. Changes are happening relatively fast (<30min) and the flow is also lower than during the steady operation.

Above 1 vol-% CO_2 concentration was observed during the start-up for a time interval of a bit less than 30 minutes. Reason for this is the same as for high temperature and pressure swings in the methanation: compressors for the recirculation and the natural gas grid compression having too fast start-up ramp. Other than this the product gas quality has found to be good during start-up and shutdown.

5 Conclusions and future work

This paper presented a simulation-aided control development study for a PtG plant. In the results we showed example simulations of plant start-up and a shutdown. The first contribution of the paper was to show the potential of using dynamic simulation in identifying control development needs in the PtG process. The second contribution was the actual development needs, namely:

- CH_4 compression start-up to minimize the amount of CO_2 to the natural gas grid and to minimize the temperature and pressure swings in the methanation reactor.
- Nitrogen purge for alkaline electrolyser.
- Effect of temperature in the alkaline electrolyser.

In the future, partial shutdown and start-up sequences will be implemented into the model, allowing for example the

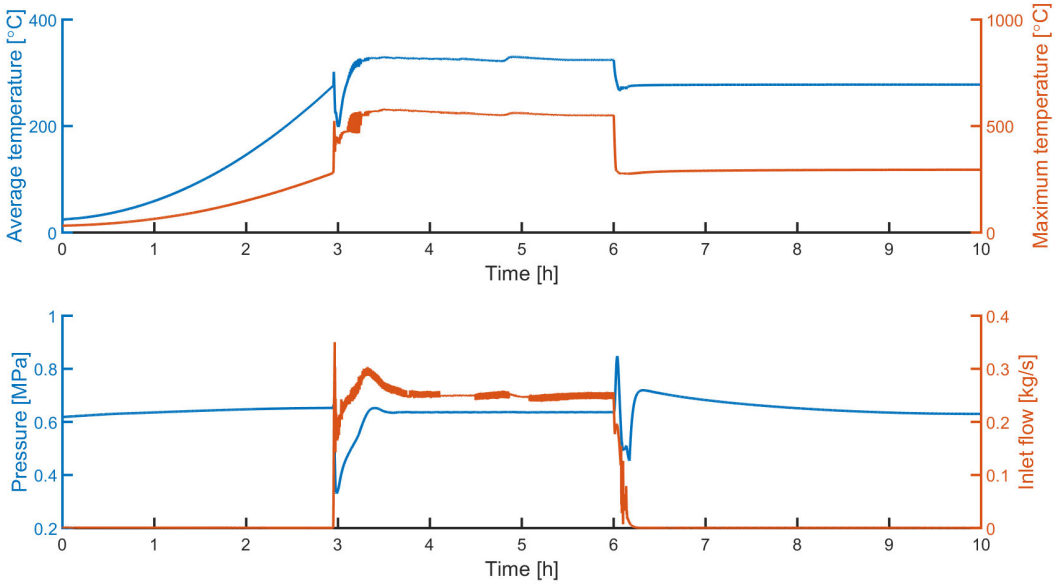


Figure 4. Changes on average and maximum temperature (upper) and pressure and inlet gas flow in the methanation reactor during start-up and shutdown sequences.

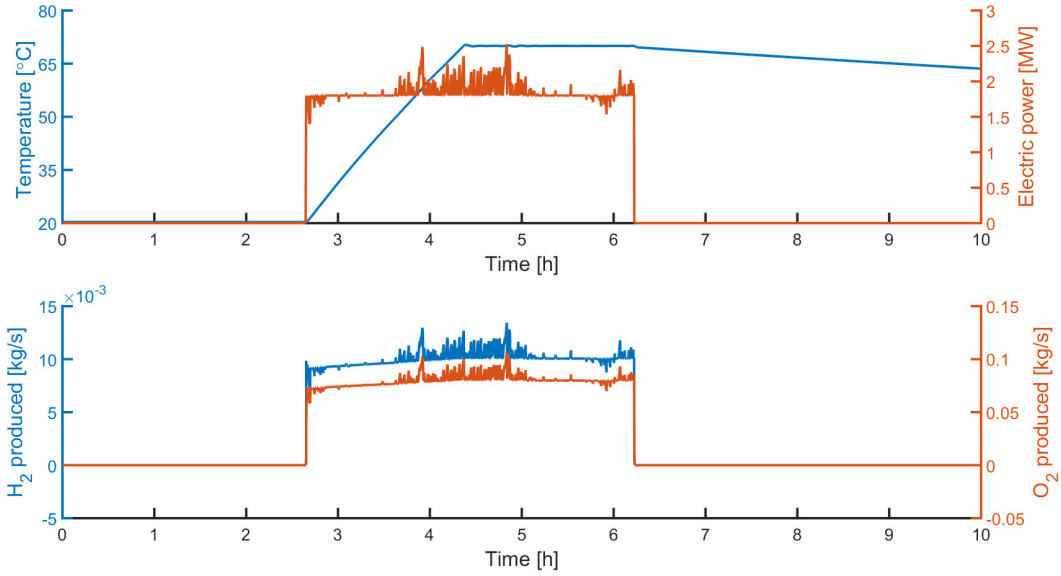


Figure 5. Temperature and inlet power changes (upper) and changes on H_2 and O_2 production (lower) during start-up and shutdown sequences of one electrolyser.

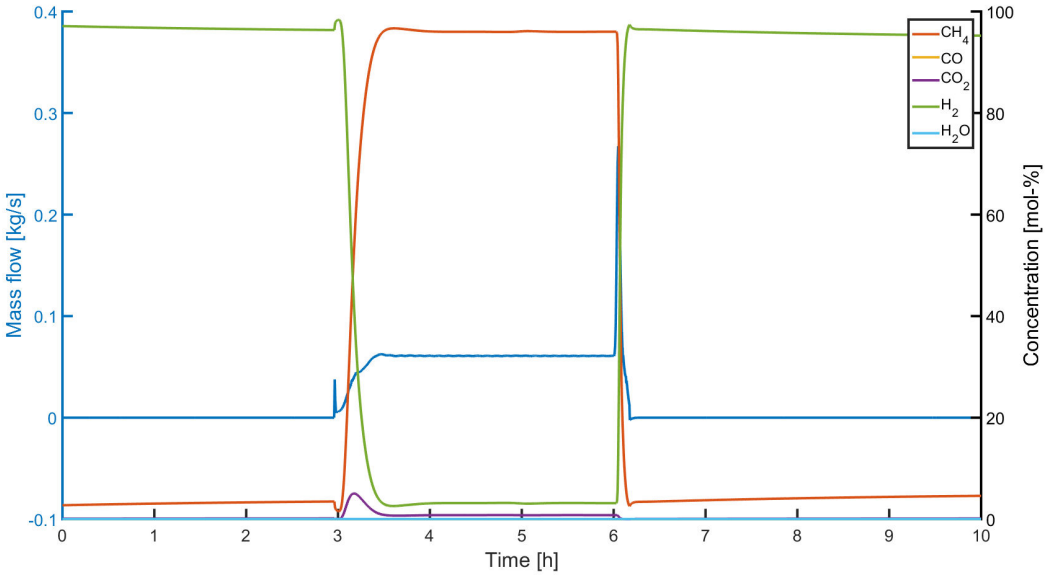


Figure 6. Mass flow and component-wise concentration changes in gas flow to natural gas grid during start-up and shutdown sequences.

methanation to be shut down when either hydrogen or carbon dioxide storage is empty. Also, optimization could be done e.g. on how to do the electrolysis start-up, with maximum electric power vs. frequency control mode used in this study. Also, heat integrations are planned to be done for the model and their effect to the start-up and shutdown logics need to be studied.

While this study presented only one case study, it can be argued that the PtG plant wide dynamic simulation aided control development approach taken here has wider applicability. Firstly, the process under study is not in any way uncommon to the chemical industry as it contains typical unit operations such as electrolysis, reactors, mixing, separations and storage. Secondly, the unit operations models utilized in the study are validated. Thirdly, as the results show, we were able to identify control development needs using the approach. Thus, it stands to reason that a similar approach is likely to be applicable and useful in other cases as well.

6 Acknowledgment

The authors gratefully acknowledge the public cofinancing of Tekes, the Finnish Funding Agency for Innovation, for the 'Neo-Carbon Energy' project under grant number 40101/14.

References

dena German Energy Agency. Power to gas system solution. opportunities, challenges and parameters on the way to marketability., 11 2015. URL http://www.powertogas.info/fileadmin/content/Downloads/Brosch%C3%BCren/dena_PowertoGas_2015_engl.pdf.

J. Eichman, K. Harrison, and M. Peters. *Novel Electrolyzer Applications: Providing More Than Just Hydrogen*. Sep 2014. doi:10.2172/1159377. URL http://www.osti.gov/scitech/servlets/purl/1159377.

European Renewable Energy Council EREC. Mapping renewable energy pathways towards 2020, 2011.

Manuel Götz, Jonathan Lefebvre, Friedemann Märs, Amy McDaniel Koch, Frank Graf, Siegfried Bajohr, Rainer Reimert, and Thomas Kolb. Renewable power-to-gas: A technological and economic review. *Renewable Energy*, 85:1371 – 1390, 2016. ISSN 0960-1481. doi:https://doi.org/10.1016/j.renene.2015.07.066. URL http://www.sciencedirect.com/science/article/pii/S0960148115301610.

Lars Jürgensen, Ehiaze Augustine Ehimen, Jens Born, and Jens Bo Holm-Nielsen. Dynamic biogas upgrading based on the sabatier process: Thermodynamic and dynamic process simulation. *Bioresource Technology*, 178:323 – 329, 2015. ISSN 0960-8524. doi:https://doi.org/10.1016/j.biortech.2014.10.069. URL http://www.sciencedirect.com/science/article/pii/S0960852414014916.

Marc Melaina, Olga Sozinova, and Michael Penev. Blending hydrogen into natural gas pipeline networks: A review of key issues, 01 2013.

Anand B. Rao and Edward S. Rubin. A technical, economic, and environmental assessment of amine-based co2 capture technology for power plant greenhouse gas control. *Environmental Science & Technology*, 36(20):4467–4475, 2002. doi:10.1021/es0158861. URL http://dx.doi.org/10.1021/es0158861. PMID: 12387425.

Jouni Savolainen, Lotta Kannari, Jari Pennanen, Matti Tähtinen, Teemu Sihvonen, Riku Pasonen, and Robert Weiss. Operation of a PtG plant under power scheduling, 03 2016.

E. Silvennoinen, M. Hänninen, K. Juslin, K. Juslin, Valtion teknillinen tutkimuskeskus, K. Porkholm, and O. Tiihonen. *The APROS Software for Process Simulation and Model Development*. Tutkimuksia (Valtion teknilli-nen tutkimuskeskus). Technical Research Centre of Finland, 1989. ISBN 9789513834630. URL https://books.google.fi/books?id=qBv9AAAACAAJ.

Matti Tähtinen, Teemu Sihvonen, Jouni Savolainen, and Robert Weiss. Interim H2 storage in power-to-x process: Dynamic unit process modelling and dynamic simulations case process and modelling principles, 03 2016.

United Natios UN. Paris agreement, 12 2015.

Robert Weiss, Jouni Savolainen, Pasi Peltoniemi, and Eero Inkeri. Optimal scheduling of a P2G plant in dynamic power heat and gas markets, 03 2016.

Øystein Ulleberg, Torgeir Nakken, and Arnaud Eté. The wind/hydrogen demonstration system at Utsira in Norway: Evaluation of system performance using operational data and updated hydrogen energy system modeling tools. *International Journal of Hydrogen Energy*, 35(5):1841 – 1852, 2010. ISSN 0360-3199. doi:https://doi.org/10.1016/j.ijhydene.2009.10.077. URL http://www.sciencedirect.com/science/article/pii/S0360319909016759.

Tao Zhou and Bruno Francois. Modeling and control design of hydrogen production process for an active hydrogen/wind hybrid power system. *International Journal of Hydrogen Energy*, 34(1):21 – 30, 2009. ISSN 0360-3199. doi:https://doi.org/10.1016/j.ijhydene.2008.10.030. URL http://www.sciencedirect.com/science/article/pii/S0360319908013141.

U Zuberbüler, M Specht, H Jachmann, S Schwarz, B Stürmer, B Feigl, and S Steiert. Project: automatic operation of a power-to-gas plant based on simulated timetables in scenarios with high shares of renewable power and the development of a smart gas grid injection concept, 03 2016.

Simulation of Particle Segregation in Fluidized Beds

Janitha C. Bandara Rajan K. Thapa Britt M.E. Moldestad Marianne S. Eikeland

Process Energy and Environmental Technology, University College of Southeast Norway, Norway
{Janitha.bandara, rajan.k.thapa, britt.moldestad, Marianne.Eikeland}@usn.org

Abstract

Fluidization technology is widely used in solid processing industry due to the high efficiency, high heat and mass transfer rate and uniform operating conditions throughout the reactor. Biomass gasification is an emerging renewable energy technology where fluidized bed reactors are more popular compared to fixed bed reactor systems due to their scalability to deliver high throughput. Fluidization of large biomass particles is difficult, and the process is therefore assisted by a bed material with higher density. The combination of different types of particles makes it challenging to predict the fluid-dynamic behavior in the reactor. Computational particle fluid dynamics simulations using the commercial software Barracuda VR were performed to study the fluidization properties for a mixture of particles with different density and size. The density ratio for the two types of particles was six, which is the typical ratio for bed material to biomass in a gasifier. The results from simulations with Barracuda VR regarding bed pressure drop and the minimum fluidization velocity, show good agreement with available experimental data. The deviation between experimental data and simulations are less than 12%. Particle segregation was clearly observed both in the simulations and in the experimental study.

Keywords: fluidization, particle density and size, pressure drop, minimum fluidization velocity, drag models, computational particle fluid dynamics

1 Introduction

Fluidized beds are extensively utilized in mechanical processes, energy technology and processing industries (Vollmari et al, 2016). The process was introduced in fluid catalytic cracking (FCC) in the petroleum industry as an alternative to thermal cracking (Horio, 2013; Winter and Schratzer, 2013).

In contrast to the fluid state reactions, solid state processing is rather challenging. Fluidization emerged as a remarkable technique due to high efficiency, low emissions and excellent heat and mass transfer characteristics (Winter and Schratzer, 2013). However, even with the knowledge over decades, the complex dynamics of the fluidization is not fully understood (Sánchez-Delgado et al, 2011).

Thermal conversion of biomass to energy is a renewable and deep-rooted process where fluidized bed combustors and gasifiers are attaining more attention compared to the fixed bed designs, due to the flexibility of scaling up along with uniform operating conditions throughout the reactor (Sau et al, 2007). A bed-material is commonly used as a fluidizing aid, because the biomass particles alone are more difficult to fluidize. Apart from assisting fluidization, the bed material can act either as a thermal energy carrier, a catalyst or both. Therefore, the presence of particles having different densities and sizes is a common in fluidized bed combustors and gasifiers since the bed material, biomass and char are fluidized together. This may lead to segregation where lighter biomass particles rise to the top of the bed while the heavier bed material moves to the bottom. However, it is said that the difference in density affects more than the difference in particle size on the tendency of segregation (Cooper and Coronella, 2005). Minimum fluidization velocity is another important parameter for gasification, because the system operates under restricted gas supply. If the mass flowrate relevant for the bubbling fluidization velocity is greater than the stoichiometric requirement for gasification, more fully oxidized products such as carbon dioxide (CO_2) could be present in the producer gas, which consequently reduces the producer gas quality. In such situations, it is necessary to bring down the minimum fluidization velocity by making necessary changes in the particle properties and the reactor geometry. Hence, it is important to study the effect of different particle mixtures on the solid flow pattern, the minimum fluidization velocity and the pressure drop behavior.

2 Fluidization

A collection of particles, such as a sandy beach, acts as a solid due to gravitational and surface forces. However, it is possible to change the solid like existence of a collection of particles by applying an upward fluid flow with a sufficient velocity to counter-weight the prior mentioned forces with the drag force. Once the static forces between the particles are overcome and all the forces are in perfect balance, the particles starts behaving like a fluid, which is known as fluidization. The minimum superficial gas velocity at which the

transition from fixed to fluidized bed occurs is referred as the minimum fluidization velocity (Horio, 2013).

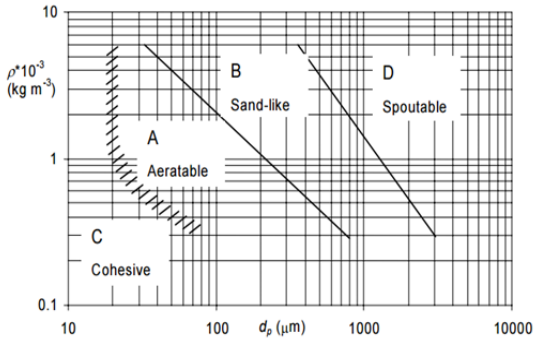


Figure 1. Geldart powder classification

Size, shape and density of the particles together with the fluid superficial velocity, characterize the behavior of fluidization where homogeneous, bubbling, slugging, spouting and turbulent are the well-known regimes. Figure 1 represents Geldart's classification of particles into the four different regimes A, B, C and D (Geldart, 1973; Fotovat et al, 2015).

According to the Geldart, type "A" particles are small with typical density is less than 1400 kg/m^3. The bed expands at the minimum fluidization velocity and the regime changes to bubbling fluidized bed as the gas velocity increases. Most of the type B materials are in the range of 40 μm to 500 μm in size with density ranging from 1400 kg/m^3 to 4000 kg/m^3. The bubbling regime starts as soon as the minimum fluidization velocity is reached. Due to fine particles and strong cohesive forces, group C powders show high resistance against fluidization. Particles with large mean size are classified as group D, which normally operates at the spouting regime. At high superficial gas velocities, all the four types of powders obtain fast and turbulent fluidization, and at further increase in the gas velocity, pneumatic transport will occur.

Pressure drop, minimum fluidization velocity and solid flow patterns are important parameters for design, optimization and operation of fluidized beds. In bubbling fluidization, the bubble activity is the main factor, which influences on the solid flow pattern inside the reactor. Particles are carried up inside the bubble, which is referred to as the bubble wake. These particles are brought to the surface of the bed, which leads to mixing of the top and bottom fractions (Sau et al., 2007). As bubbles move upwards, the void leaving behind immediately below the bubble is filled with surrounding particles. Whenever multiple particle species are involved, heavier particles tend to accumulate at the bottom and lighter at the bed surface, which are referred as 'jetsam' and 'flotsam' respectively (Cooper and Coronella, 2005). Solid flow patterns affect the characteristics such as rate of heat and mass transfer, reaction intensity, particle attrition, mixing, segregation and internal corrosion. In addition to the particle and gas properties, bed geometry, aspect ratio and gas distribution make a considerable influence on the fluidization parameters.

Experimental investigations at industrial scale fluidization involve high costs, while it is not easy to scale up from lab scale to industrial scale. Hence, detailed analysis of the fluidization process via simulations provides a better understanding at reduced costs and time (Vollmari et al, 2016).

Computational Fluid Dynamics (CFD) has become a powerful tool, and different software such as Barracuda VR, FLUENT, OpenFOAM and MFIX have been developed for CFD simulations. The objective of this paper is to analyze the effect of the presence of particles with different densities and sizes in the fluidized bed, on the pressure drop and minimum fluidization velocity, using the commercial CPFD software Barracuda VR. However, as some correlations used in CFD packages are empirical or semi-empirical, it is important to validate the results against experimental data performed on a fluidized bed with the same size and design (Taghipour et al, 2005). Hence, the simulation results are validated against data from experimental studies performed by Thapa et al (2011).

3 Experimental Setup

Thapa et al. (2011) have carried out experimental investigations of fluidization using particles with different densities and sizes. The experimental rig was an 84mm diameter cylindrical column with a homogeneous air distributor at the bottom. Nine pressure measuring points have been used which were equally spaced with 100 mm along its height. Figure 2 illustrates the arrangements of the experimental rig.

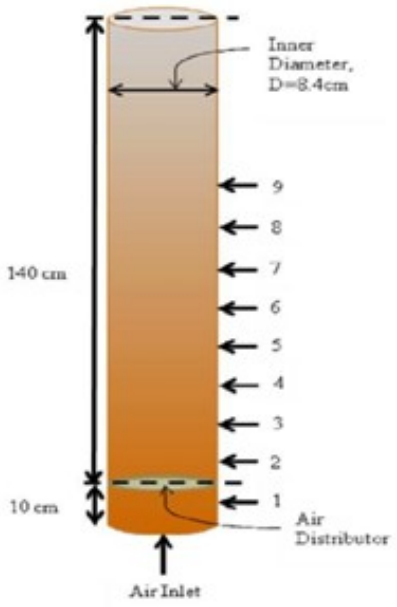


Figure 2. Geldart powder classification

In a real gasification system, the densities of the bed material and char/wood particles are approximately 2500 kg/m^3 and 400-500 kg/m^3 respectively. Here, ZrO_2 and plastic beads were used to represent the bed material and biomass respectively, keeping the density ratio at a value of six, which is about the same as for a real gasification system.

4 Simulation Setup

Two fluid model (TFM) and Discrete particle/element model (DPM/DEM) are the two basic approaches, which are also referred to as the Eulerian-Eulerian and the Eulerian-Lagrangian models respectively. TFM considers that both the fluid and the particle phases are fully interpenetrating continua, where the solid phase properties are calculated based on the kinetic theory of granular flow (Jayarathna, 2014).

In the DPM simulations, the fluid phase is treated as a continuous medium whereas the particles are considered as individual components. Using DPM, the equations are solved for each particle. This approach needs the collision models to address the non-ideal interaction between particles. However, the DPM approach is computationally expensive, and it is only possible to handle a limited number of particles.

Computational particle fluid dynamics (CPFD) is an extension of CFD, which facilitates the multi-phase systems such as fluidization. Barracuda VR is specially developed for multiphase CPFD simulations, which utilizes the theory of multiphase Particle in Cell (MP – PIC). In the PIC approach, a collection of real particles, which is referred as a 'computational particle', is performed in the Lagrangian particle modeling while the fluid phase is treated as a continuum (Andrews and O'Rourke, 1996).

The computational setup is presented in Figure 3. Air exit at the top of the column was set as a pressure boundary with atmospheric conditions and with no particle exit. Stepwise increase of the air velocity at the inlet flow boundary (bottom of the geometry) was performed in order to calculate the minimum fluidization velocity. Both pressure and temperature were defined at the inlet flow boundary, which allows the system to calculate the mass flow rate of the gas. However, this inlet pressure applies only for the initial time step and the system itself calculates the required pressure according to the bed pressure drop. The time step was set to 0.001 seconds for the simulations and whenever the user defined time step is too high, Barracuda VR re-calculates it. This guarantees the stability of the simulation.

Simulations were carried out for both the individual particle types and for mixtures of the two types of particles. The effect on the minimum fluidization velocity was observed. Table 1 summarizes the rest of the simulation parameters.

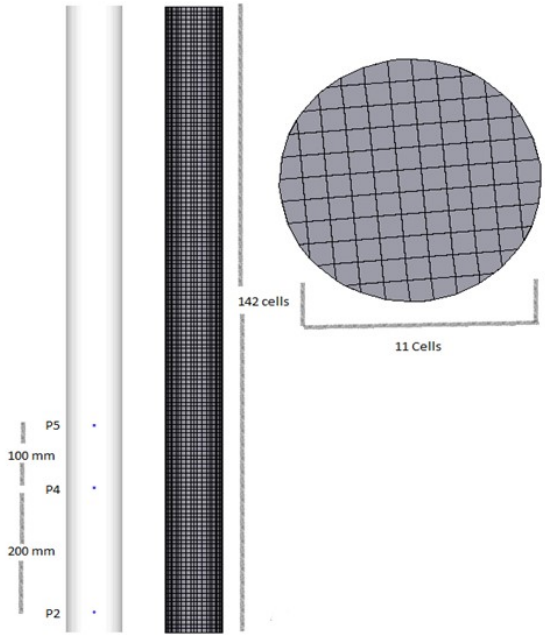


Figure 3. Computational grid for the simulation and pressure transient data points

As illustrated in Figure 3, the pressure monitor points P2, P4 and P5 were included. Uniform meshing and 17061 cells were used for all the simulations. The system was simulated for 5 seconds at each velocity step and the pressure, averaged over the last second of each time step, was used in the data analysis. The pressure drop was calculated as the pressure drop per unit height (mbar/m) in order to minimize the errors. The pressure drop gradient as a function of the superficial gas velocity for both experimental data and simulation results, were plotted in the same chart to compare the results.

Table 1. Simulation parameters

Parameter	*Value*
ZrO_2 density	5850kg/m^3
Plastic density	964kg/m^3
ZrO_2 particle size	709 µm
Plastic particle size	3500 µm
Bed height	340 mm
Bed width	84 mm
Temperature	300 K
Sphericity	1
Maximum packing	0.6
Max. Momentum redirection from collisions	40%
Normal to wall momentum retention	0.3
Tangent to wall momentum retention	0.99
Stress model parameters	Default

5 Results and Discussion

A real gasification/combustion system is a reaction volume with non-isothermal conditions with a wide range of particle sizes, shapes and densities. A fluidized bed gasifier includes wood chips (1-5 cm), reacted char (larger particle range) and bed material (400 to 600 microns). Large plastics beads with a wide shape distribution are used in experiments and simulations providing a good approximation to the wood particles. A density ratio of six between the two particle types is used to get a good agreement between the cold flow model and the actual gasification model. The use of air at ambient temperature in contrast to the actual system of high temperature air or steam is another challenge when predicting the real world system. However, cold bed data may provide an overview to get a good understanding of fluidization of particle mixtures having different densities and particle sizes. The study is an initial step in simulation of a complex fluidized bed gasification process.

5.1 Minimum Fluidization Velocity and Bed Pressure Drop

The simulations were carried out for pure ZrO_2 particles, pure plastic beads and with mixtures of 10% and 20% plastic beads with ZrO_2 as the rest. The pressure drop gradient over the bed height was calculated using equation (1).

$$\frac{\Delta P}{H} = \frac{P_4 - P_2}{200} \tag{1}$$

The minimum fluidization velocity is determined as the velocity where the increasing pressure drop starts to drop creating a maximum. However, the minimum fluidization velocity lies slightly below that value due to initial packed bed resistance. Hence, the best way to determine the minimum fluidization velocity is to carry out the experiments from high velocities to lower velocities as well, and get the average values for the two cases. Wen-Yu, Ergun drag model was used in the Barracuda VR simulations and it is available as an inbuilt model.

Experimental data for the pure plastic beads and simulation results based on the Wen-Yu-Ergun drag model is illustrated in Figure 4. The Wen-Yu-Ergun model is a combined model for the drag force, which was proposed by Gidaspow. The combined model selects the interface momentum transfer coefficient (K_{sg}) by either Wen-Yu or Ergun depending upon the bed void fraction. If the gas volume fraction is greater than 0.8, the Wen-Yu correlation is selected. The Wen-Yu correlation is expressed in equation (2).

$$K_{sg} = \frac{3}{4}\frac{C_D \rho_g \varepsilon_g (1-\varepsilon_g)(u_s - u_g)}{d_p}\,\varepsilon_g^{-2.65} \tag{2}$$

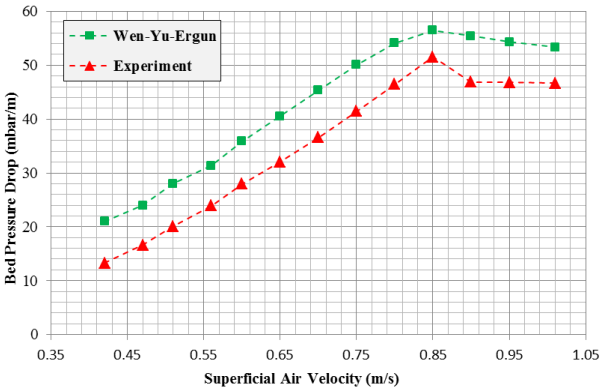


Figure 4. Simulation and experimental data for plastic beads

Where CD is given by:

$$C_D = \frac{24}{\varepsilon_g Re_s}\left[1 + 0.15\left(\varepsilon_g Re_s\right)^{0.687}\right] \tag{3}$$

Otherwise, when the void fraction is lower than 0.8, the Ergun correlation is selected as in equation (4):

$$K_{sg} = 150\frac{\mu_g(1-\varepsilon_g)^2}{\varphi^2 {d_p}^2 \varepsilon_g} + 1.75\frac{\rho_g(u_g - u_s)(1-\varepsilon_g)}{\varphi d_p} \tag{4}$$

Where u_g is the superficial gas velocity, μ_g is the gas viscosity, ρ_g is the density of the gas, ρ_s is the density of particles, d_p is the mean particle diameter, φ is the particle sphericity, Re_s is the particle Reynolds number and ε_g is the volume fraction of the gas.

Experimental results for plastic beads show that the minimum fluidization velocity and maximum bed pressure drop are 0.85 m/s and 50 mbar/m respectively. The pressure drop from the simulations is higher compared to the experimental data. The pressure drop gradient at minimum fluidization velocity is measured to be 52 mbar/m in the experiments and calculated as 56 mbar/m in the simulations. This deviation are mainly due to the uncertainty of bed void fraction data, particle sphericity and particle size distribution. Bed voidage is a strong function of particle size distribution and sphericity. Simulations with increased void fraction illustrated that the bed pressure drop reached the experimental values without changing the minimum fluidization velocity. Further, the pressure tapping points are at the wall of the experimental rig and in the real situation, there is a tendency of air to escape through loosely packed regions near the wall.

However, the minimum fluidization velocity is the same for the experiments and the simulations, having a value of 0.85 m/s. The simulation results for the bed pressure drop is higher than the experimental data also after the fluidization regime is reached. Parameters of the particle stress model affect considerably in the fluidization regime while it has a negligible effect in the packed bed region. As the main objective for this study

was to analyze the minimum fluidization velocity and particle segregation, default values of the particle stress model were used. This can be a reason for the deviation in pressure drop in the fluidization regime.

The behavior of pure ZrO_2 particles is plotted in Figure 5. The pressure drop from the simulations is slightly higher than the experimental data. The simulated bed pressure drop at minimum fluidization is 350 mbar/m, which gives a deviation of about 6% compared to the experimental data of 330 mbar/m. Minimum fluidization velocities are 0.67m/s and 0.7m/s in the experiment and simulation respectively. The particle size distribution is an important factor for both the minimum fluidization velocity and the bed pressure drop. When the particle size distribution is wide, the void fraction in the bed becomes lower. However, the particle size distribution for ZrO_2 in the experimental study has not been reported, and only the mean diameter of 709 µm was available. Therefore, using uniform size particles could be one of the reasons for the deviation in the results.

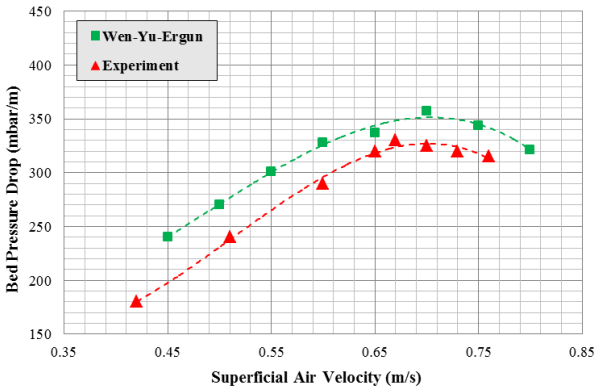


Figure 5. Simulation and experimental data for ZrO_2

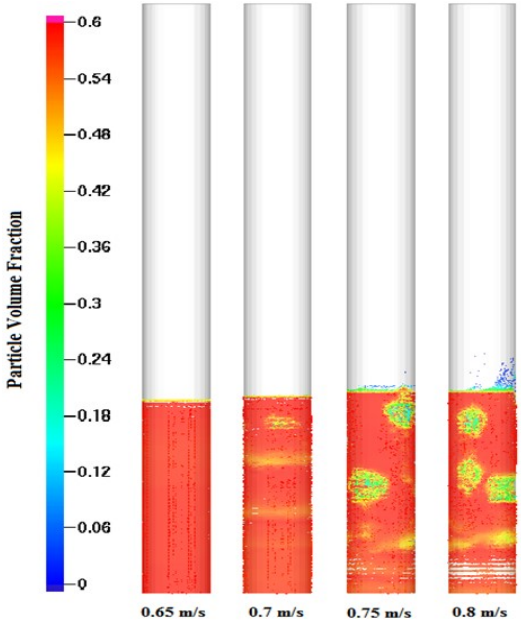


Figure 6. Transition from fixed to fluidized bed of ZrO_2

A graphical display of transition from fixed bed to fluidized bed followed by bubble formation for ZrO2 particles is captured in Figure 6. There was a high tendency of bubble formation and bubble rise near the walls.

The simulation results for pure particles of plastic and ZrO_2 agree well with the experimental results. Simulations using different drag models were performed, and Wen-Yu-Ergun was found to have the best agreement with the experimental data. Therefore, simulations for particle mixtures of plastic and ZrO_2 were performed with this drag model. Figure 7 and Figure 8 depict the superficial air velocity versus bed pressure drop for particle mixtures of 10% and 20% plastic beads respectively.

The simulation results of both the mixtures show similar behaviors. The bed pressure drop is higher than the experimental data in the fixed bed regime, while the gradient sharply drops down close to the minimum fluidization velocity. In contrast to the case of pure particles, it is difficult to observe a sharp change in pressure drop gradient at the state of minimum fluidization. Typically, when two particles with different minimum fluidization velocities are mixed, local fluidization behavior of one component (with lower minimum fluidization velocity) can be observed before the total bed starts to fluidize. This phenomenon breaks the linear relationship of air velocity to pressure drop in the fixed bed regime.

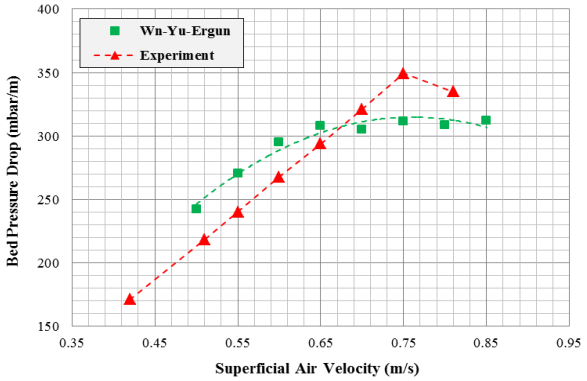


Figure 7. Simulation results 10% plastic and ZrO_2

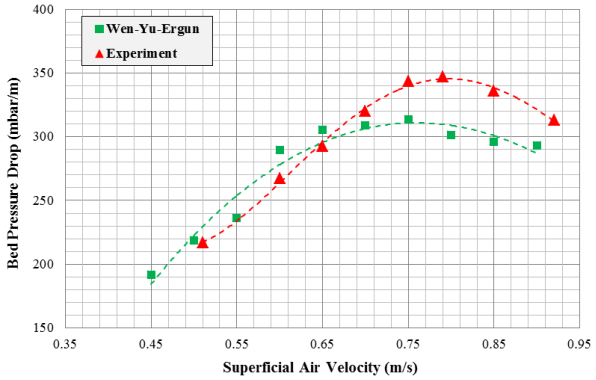


Figure 8. Simulation results 20% plastic and ZrO_2

In the simulations, the pressure drop tends to be about constant when the gas velocity changes from 0.65 to 0.75 m/s. After reaching the minimum fluidization velocity, the pressure gradient starts to decrease. Both mixtures have high fraction of ZrO_2 and that could be one reason for the steady pressure drop when the velocity exceed 0.65m/s, which is approximately the minimum fluidization velocity of ZrO_2. The experimental minimum fluidization velocities were 0.75 m/s and 0.77 m/s for the mixtures with 10% and 20% plastic particles, respectively. In the simulations, the minimum fluidization velocity was 0.75m/s for both the mixtures. It is rather difficult to pick a sharp turning point for minimum fluidization from the simulation results. However, the simulation result follows the trend of the experimental data.

5.2 Particle Segregation

The distribution of ZrO_2 particles and plastic beads in the bed as a function of time is given in Figure 9. The figure clearly depicts that the lower density particles moves upward while the higher density fraction accumulates at the bottom. Even though plastic beads have significantly higher particle size compared to ZrO_2, the lower density makes it move up to the surface. Therefore, it is vital to know the time taken for complete segregation. Once the low density particles have moved to the bed surface, the bed properties are changed from the initial conditions and this may lead to fluidization of only one of the segregated layers. The segregation of particles was also observed in the experimental study of Thapa et al (2011).

When simulating mixtures with different particle sizes in Barracuda VR, it is recommended to activate the blended acceleration model (BAM) which averages the different velocities caused by the size differences.

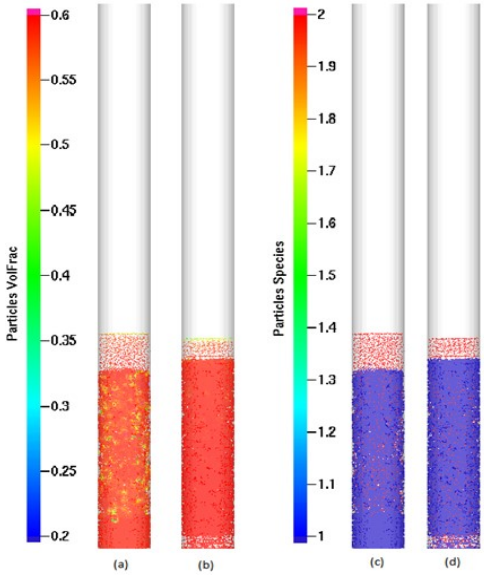


Figure 9. Particle segregation of a mixture of ZrO_2 and 10% plastic beads, a) and c) 15s after fluidization, b) and d) 5s after fluidization

6 Conclusions

The results from simulations with Barracuda VR show good agreement with available experimental data. The deviation between experimental data and simulations are less than 12% for both the bed pressure drop and the minimum fluidization velocity.

The selection of drag model is crucial for the simulation results, and therefore the model should be carefully chosen. However, it is challenging to predict the suitability of a particular drag model for a particular system and it is recommended to get the model validated against experimental data. The deviation between experimental data and simulations were less than 12%. Insufficient information about particle size distribution for the ZrO_2, together with uncertainty about particle sphericity and the particle volume fraction can influence on the simulation results. The large size of plastic beads was a constraint in order to use a fine mesh in the simulations.

The users can define their own drag models in Barracuda VR, and are allowed to change the particle normal stress model parameters. However, the system can be further optimized by changing the drag model, the particle stress model parameters and the wall interaction coefficients.

Particle separation was clearly observed and it should be further analyzed with different air velocities and with different mixtures of particles. More analysis is possible with varied particle size to get the optimized size ratio for the particle mixtures to avoid segregation.

Acknowledgements

The authors like to thank Mr. Chameera Jayarathna and Mr. Amila Chandra for their support in handling Barracuda VR and SolidWorks.

References

M.J. Andrews and P. J. O'Rourke . The multiphase particle-in-cell (MP-PIC) method for dense particulate flows. *International Journal of Multiphase Flow,* 22(2): 379-402, 1996.

S. Cooper and C. J. Coronella. CFD simulations of particle mixing in a binary fluidized bed. *Powder Technology,* 151(1): 27-36, 2005.

F. Fotovat, R. Ansart, M. Hemati, O. Simonin and J. Chaouki. Sand-assisted fluidization of large cylindrical and spherical biomass particles: Experiments and simulation. *Chemical Engineering Science,* 126(Supplement C): 543-559, 2015.

D. Geldart. Types of gas fluidization. *Powder Technology,* **7**(5): 285-292, 1973.

M. Horio. 1 - Overview of fluidization science and fluidized bed technologies A2 - Scala, Fabrizio. *Fluidized Bed Technologies for Near-Zero Emission Combustion and Gasification*, Woodhead Publishing**:** 3-41, 2013.

C.K. Jayarathna, B. M. Halvorsen and L.-A. Tokheim. Experimental and Theoretical Study of Minimum Fluidization Velocity and Void Fraction of a Limestone Based CO_2 Sorbent. *Energy Procedia,* 63(Supplement C): 1432-1445, 2014.

S. Sánchez-Delgado, J. A. Almendros-Ibáñez, N. García-Hernando and D. Santana. On the minimum fluidization velocity in 2D fluidized beds. *Powder Technology,* 207(1): 145-153, 2011.

D.C. Sau, S. Mohanty and K. C. Biswal. Minimum fluidization velocities and maximum bed pressure drops for gas–solid tapered fluidized beds. *Chemical Engineering Journal,* 132(1): 151-157, 2007.

F. Taghipour, N. Ellis and C. Wong. Experimental and computational study of gas–solid fluidized bed hydrodynamics. *Chemical Engineering Science,* 60(24): 6857-6867, 2005.

R.K. Thapa, C. Rautenbach and B. M. Halvorsen. Investigation of flow behavior in biomass gasifier using electrical capacitance tomography (ECT) and pressure sensors. *International Conference on Polygeneration Strategies*. F. M. Hofbauer. Vienna University of Technology, Austrian National Library**:** 97-106, 2011.

K. Vollmari, R. Jasevičius and H. Kruggel-Emden. Experimental and numerical study of fluidization and pressure drop of spherical and non-spherical particles in a model scale fluidized bed. *Powder Technology,* 291(Supplement C): 506-521, 2016.

F. Winter and B. Schratzer. 23 - Applications of fluidized bed technology in processes other than combustion and gasification A2 - Scala, Fabrizio. *Fluidized Bed Technologies for Near-Zero Emission Combustion and Gasification*, Woodhead Publishing**:** 1005-1033, 2013.

Dynamic Model of an Ammonia Synthesis Reactor based on Open Information

Asanthi Jinasena Bernt Lie Bjørn Glemmestad

Faculty of Technology, University College of Southeast Norway, Norway,
{asanthi.jinasena,Bernt.Lie}@usn.no

Abstract

Ammonia is a widely used chemical, hence the ammonia manufacturing process has become a standard case study in the scientific community. In the field of mathematical modeling of the dynamics of ammonia synthesis reactors, there is a lack of complete and well documented models. Therefore, the main aim of this work is to develop a complete and well documented mathematical model for observing the dynamic behavior of an industrial ammonia synthesis reactor system. The model is complete enough to satisfactorily reproduce the oscillatory behavior of the temperature of the reactor.

Keywords: modeling, ammonia, reactor, dynamic, simulation

1 Introduction

The control of the ammonia synthesis reactor is an interesting topic in the industrial and scientific community, because of the importance and the dynamics of it. Mathematical modeling of the ammonia synthesis loop is a common strategy for understanding and controlling these dynamics. Most of the studies are focused on steady state operation. Simulation of ammonia synthesis reactors for design, optimization (Baddour et al., 1965; Murase et al., 1970; Singh, 1975) and control (Shah, 1967; Singh and Saraf, 1979) has been reported since the late 1960s. However, studies on reactor instability started a few years earlier (van Heerden, 1953). A few studies have been done on dynamic modeling of ammonia synthesis reactors. However, most available models are incomplete in information: missing parameter values, missing or incorrect units, missing expressions for reaction rate due to confidentiality, inconsistent thermodynamics and missing operating conditions. The main objective of this study is therefore to compile a complete, well–documented and easily accessible dynamic model purely based on information available through open publications. The model is used to reproduce the oscillatory behavior of temperature which has been reported especially on manually controlled industrial reactors (Naess et al., 1993; Morud and Skogestad, 1993; Morud, 1995; Morud and Skogestad, 1998; Rabchuk et al., 2014; Rabchuk, 2014).

Considering the few dynamic models that have been reported, Naess et al. (1993) developed a model for optimization and control of the ammonia synthesis process based on an incident of an ammonia synthesis plant in Germany. The simulations were verified using the plant data. For the same incident, Morud (1995), Morud and Skogestad (1993, 1998) analyzed the instability through a dynamic model to reproduce the behavior of rapid temperature oscillations observed in the industrial ammonia reactor system by stepping down the reactor pressure. A linear dynamic analysis was done on the model. It has been shown that the cause of the limit cycle behavior of the reactor was positive temperature feedback from the heat exchanger and a non-minimum phase behavior of the temperature response. A feedback controller is suggested to control this behavior (Morud and Skogestad, 1998).

Rabchuk (2014) and Rabchuk et al. (2014) have developed a dynamic model for testing the stability of an industrial ammonia synthesis reactor. The system consisted of a catalytic bed ammonia synthesis reactor and a heat exchanger and the oscillations were obtained by stepping down the feed temperature. A stability analysis was also done for selected process parameters (Rabchuk, 2014).

This paper consists of a detailed model description in Section 2, including the assumptions, the topology and the descriptive model development for the reactor and the heat exchanger. This is followed by the simulation results and discussion with a comparison with previous work in Section 3. All the values and units for the used parameters and operating conditions are included in the Appendix.

2 Mathematical Model

The Haber–Bosch process is used to produce ammonia from the following reaction using an iron based catalyst,

$$\mathrm{N_{2(g)}} + 3\,\mathrm{H_{2(g)}} \overset{\mathrm{Fe}}{\rightleftharpoons} 2\,\mathrm{NH_{3(g)}}\,. \tag{1}$$

Argon (Ar) is also present as an inert gas. The ammonia synthesis process includes catalytic bed reactors for ammonia formation with heat exchangers, where the product gas streams are cooled by the feed gas streams. A simplified diagram of the reactor configuration is shown in Figure 1. The reactor is considered to have one fixed catalytic bed and no bypass or intermediate cooling gas streams for simplicity. The heat exchanger is considered to be a simple counter current heat exchanger.

In this system, $\dot{m}_i$ and $\dot{m}_o$ are the inlet and outlet mass flow rates of the system, respectively. T_i is the temperature

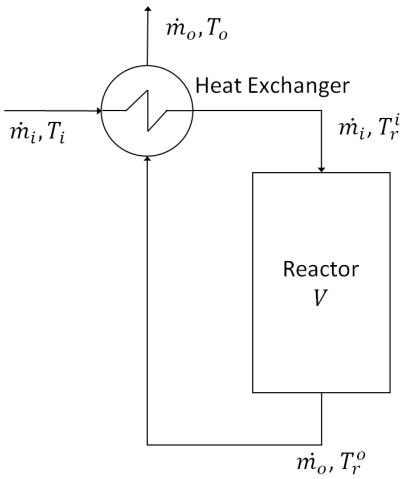


Figure 1. A simplified reactor configuration.

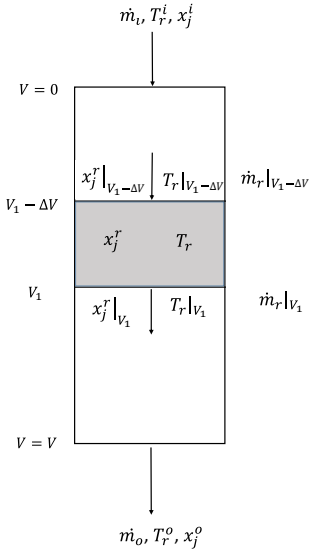


Figure 2. A schematic diagram of the distributed reactor model.

of the inlet flow of the heat exchanger. T_o is the temperature of the flow out of the heat exchanger, T_r^i is the temperature of the reactor inlet and T_r^o is the temperature of the reactor outlet. The volume of the reactor is denoted by V. The input $\dot{m}_i$ and the set point to the reactor pressure controller can be manipulated. T_i and the inlet mole fractions of various species $\left(x_j^i\right)$ are considered as disturbances to the system. The temperature of the reactor T_r is the output of interest.

2.1 Assumptions

The following assumptions are used:

- The model is one-dimensional, i.e. the temperature and molar gradients only vary in the axial direction.
- The Temkin–Pyzhev reaction rate expression is valid for the system (Murase et al., 1970; Morud and Skogestad, 1998; Froment et al., 2010).
- The discretized reactor volume compartments are well mixed.
- No heat or mass diffusion in the system.
- Individual gases and gas mixture behave as ideal gas.
- The catalyst activity is uniform throughout the reactor.
- The heat transfer coefficient, heat of reaction and heat capacities are constants.
- Reactor pressure is controlled perfectly.

2.2 Development of Model

2.2.1 Material Balance

The pressure inside the reactor (p) is considered to be constant. The schematics shown in Figure 2 depicts the distributed reactor model. For volume compartment V_1, the mole balance equation can be written as shown in Eq. 2.

$$\frac{d}{dt} n_j^r\big|_{V_1} = \dot{n}_j^r\big|_{V_1-\Delta V} - \dot{n}_j^r\big|_{V_1} + \dot{n}_j^{r,g}\Big|_{V_1} \tag{2}$$

Here, n_j^r is the number of moles inside the reactor compartment at a given time t, $\dot{n}_j^r$ is the rate of moles leaving the reactor compartment and $\dot{n}_j^{r,g}$ is the rate of generation of moles inside the reactor compartment. Superscript r denotes the reactor and subscript j denotes the particular specie, where $j \in (\mathrm{H_2}, \mathrm{N_2}, \mathrm{NH_3}, \mathrm{Ar})$. The rate of generation can be expressed using the reaction rate r, stoichiometric matrix ν and the catalyst mass m_c in the reactor volume,

$$\dot{n}_j^{r,g}\Big|_{V_1} = \nu_j\, r m_c|_{V_1} = \nu_j\, r|_{V_1} m_c \frac{\Delta V}{V}, \tag{3}$$

where ν = [-3 -1 2 0]. ΔV is the volume of a reactor compartment. The rate of reaction (rate of Nitrogen consumption per unit catalyst mass) can be found using the Temkin–Pyzhev equation (Murase et al., 1970; Morud and Skogestad, 1998).

$$r|_{V_1} = \frac{f}{\rho_c}\left(k_+ \frac{p_{N_2} {p_{H_2}}^{1.5}}{p_{NH_3}} - k_- \frac{p_{NH_3}}{{p_{H_2}}^{1.5}}\right)\Bigg|_{V_1} \tag{4}$$

where f is the catalyst activity factor, ρ_c is the packing density of the catalyst, k_+ and k_- are the rate constants of the forward and reverse reactions, respectively, and p_j denotes the partial pressure of the species in the reactor compartment. Using the Gibbs Free Energy approach at a constant temperature, the reverse reaction rate can be expressed as follows (Froment et al., 2010),

$$k_- = k_0^- \exp\left(-\frac{E_-}{RT}\right). \tag{5}$$

Using the chemical kinetics of the reaction, k_+ can be found from the equilibrium constant, K_p,

$$k_+ = k_- K_p. \tag{6}$$

The value for K_p can be computed using available correlations. The Gillespie–Beattie correlation is selected as the most suitable correlation for this system (Gillespie and Beattie, 1930). The correlation is

$$K_p^{GB} = K_p^{GB*}\ 10^{\alpha \cdot p^*}, \tag{7}$$

where K_p^{GB} is the Gillespie–Beattie equilibrium constant. The pressure correction coefficient α is given as

$$\alpha = \frac{0.1191849}{T_r|_{V_1}} + \frac{91.87212}{T_r|_{V_1}^2} + \frac{25122730}{T_r|_{V_1}^4}, \tag{8}$$

where T_r is the temperature of the reactor compartment. The value of the K_p^{GB*} can be computed from Eq. 9.

$$\begin{aligned}\log K_p^{GB*} = &-2.69112 \log T_r|_{V_1} - 5.51926 \times 10^{-5}\ T_r|_{V_1} \\ &+ 1.84886 \times 10^{-7}\ T_r|_{V_1}^2 + \frac{2001.6}{T_r|_{V_1}} \\ &+ 2.6899\end{aligned} \tag{9}$$

The dimensionless pressure p^* is

$$p^* = \frac{p}{p^\sigma}, \tag{10}$$

where p^σ is the atmospheric pressure in the given pressure unit. The relationship between the two rate coefficients is given by Eq. 11.

$$K_p = K_p^{GB2} \tag{11}$$

Temperature T_r at reactor compartment V_1 can be found by rearranging the ideal gas law to express the temperature as shown in Eq. 12.

$$T_r|_{V_1} = \frac{p \cdot \Delta V}{n_r|_{V_1} R} \varepsilon \tag{12}$$

Here, ε is the void fraction of the catalyst and n_r is the total number of moles in the reactor volume where

$$n_r = \sum_j n_j^r, \tag{13}$$

and R is the universal gas constant.

2.2.2 Energy Balance

The energy balance equation for volume compartment V_1 is

$$\begin{aligned}\frac{d}{dt}\left(H - pV\right)|_{V_1} = &\dot{H}\big|_{V_1-\Delta V} - \dot{H}\big|_{V_1} \\ &+ \dot{Q}\big|_{V_1} + \dot{W}\big|_{V_1}.\end{aligned} \tag{14}$$

Assuming no heat flow $\dot{Q}$, no shaft work $\dot{W}$ and constant pressure, the Eq. 14 can be simplified into

$$\frac{d}{dt} H|_{V_1} = \dot{H}\big|_{V_1-\Delta V} - \dot{H}\big|_{V_1}, \tag{15}$$

where H is the enthalpy of the reactor volume at a given time t, $\dot{H}$ is the rate of enthalpy of the flow into/out of the reactor volume. H can be written using the enthalpies of individual components of the mixture,

$$H|_{V_1} = \sum_j n_j^r \tilde{H}_j \Big|_{V_1} + m_c \hat{H}_c\big|_{V_1} \tag{16}$$

where $\tilde{H}_j$ is the molar enthalpy of pure gas $j \in (\mathrm{H_2}, \mathrm{N_2}, \mathrm{NH_3}, \mathrm{Ar})$ and $\hat{H}_c$ is the specific enthalpy of the catalyst. Similarly $\dot{H}$ is

$$\dot{H}\big|_{V_1} = \sum_j \dot{n}_j^r \tilde{H}_j \Big|_{V_1}. \tag{17}$$

Using Eqs. 2, 16 and 17, Eq. 15 can be developed as follows,

$$\begin{aligned}\sum_j n_j^r \frac{d\tilde{H}_j}{dt}\Big|_{V_1} + m_c \frac{d\hat{H}_c}{dt}\Big|_{V_1} = &-\sum_j \dot{n}_j^{r,g} \tilde{H}_j \Big|_{V_1} \\ &+ \sum_j \dot{n}_j^r\big|_{V_1-\Delta V} \left(\tilde{H}_j\big|_{V_1-\Delta V} - \tilde{H}_j\big|_{V_1}\right).\end{aligned} \tag{18}$$

With the use of following approximations,

$$d\tilde{H}_j \approx \tilde{c}_{p,j} dT \tag{19}$$

$$\tilde{H}_1 - \tilde{H}_2 \approx \bar{\tilde{c}}_p (T_1 - T_2) \tag{20}$$

where $\tilde{c}_{p,j}$ is molar heat capacity of each gas and $\bar{\tilde{c}}_p$ is the average molar heat capacity of the gas mixture, the model can be simplified into

$$\begin{aligned}C_p \frac{dT_r}{dt}\Big|_{V_1} = &\dot{n}_r \bar{\tilde{c}}_p\big|_{V_1-\Delta V} \left(T_r|_{V_1-\Delta V} - T_r|_{V_1}\right) \\ &- \Delta\tilde{H}_r r m_c\big|_{V_1},\end{aligned} \tag{21}$$

where $\Delta\tilde{H}_r$ is the heat of reaction. Here, C_p is the heat capacity of the reactor compartment,

$$C_p = \sum_j n_j^r \tilde{c}_{p,j} + m_c \hat{c}_{p,c}, \tag{22}$$

where $\hat{c}_{p,c}$ is the specific heat capacity of the catalyst.

Taking the time derivative of ideal gas law with constant pressure and then substituting the expression in Eq. 21 for the term $\frac{dT_r}{dt}$ will lead to the Eq. 23,

$$\begin{aligned}\dot{n}_r|_{V_1} = &\dot{n}_r|_{V_1-\Delta V} + \dot{n}^{r,g}|_{V_1} + \frac{n_r}{T_r C_p}\Big|_{V_1} \Big[\dot{n}_r \bar{\tilde{c}}_p\big|_{V_1-\Delta V} \\ &\left(T_r|_{V_1-\Delta V} - T_r|_{V_1}\right) - \Delta\tilde{H}_r r m_c\big|_{V_1}\Big]\end{aligned} \tag{23}$$

which can be re-arranged into

$$T_r C_p \dot{n}^{r,g}\big|_{V_1} - \Delta\tilde{H}_r n_r r m_c\big|_{V_1} = T_r C_p \dot{n}_r\big|_{V_1} - \dot{n}_r|_{V_1-\Delta V}\left[T_r C_p\big|_{V_1} + \bar{\tilde{c}}_p\big|_{V_1-\Delta V} n_r|_{V_1} \left(T_r|_{V_1-\Delta V} - T_r|_{V_1}\right)\right]. \quad (24)$$

This can be written in matrix form,

$$b = A \cdot \dot{n}_r \quad (25)$$

where $A \in \mathbf{R}^{N\times N}$ and $\dot{n}_r, b \in \mathbf{R}^{N\times 1}$. Here N is the number of reactor compartments in the reactor. If all compartments are considered to have equal volumes of ΔV with $N = \frac{V_r}{\Delta V}$, then

$$b_1 = T_r C_p \dot{n}^{r,g}\big|_{\Delta V} - \Delta\tilde{H}_r n_r r m_c\big|_{\Delta V} + \dot{n}_r|_0 \left[T_r C_p\big|_{\Delta V} + \dot{n}_r \bar{\tilde{c}}_p\big|_0 n_r|_{\Delta V}\left(T_r|_0 - T_r|_{\Delta V}\right)\right] \quad (26)$$

$$b_i = T_r C_p \dot{n}^{r,g}\big|_{i\Delta V} - \Delta\tilde{H}_r n_r r m_c\big|_{i\Delta V}, \quad i \in \{2,3,...,N\} \quad (27)$$

and,

$$A_{i,i} = T_r C_p\big|_{i\Delta V}, \quad i \in \{1,2,...,N\} \quad (28)$$

$$A_{i,i-1} = -T_r C_p\big|_{i\Delta V} - \bar{\tilde{c}}_p\big|_{(i-1)\Delta V} n_r|_{i\Delta V} \left(T_r|_{(i-1)\Delta V} - T_r|_{i\Delta V}\right), i \in \{2,3,...,N\}. \quad (29)$$

Solving Eq. 25 gives $\dot{n}_r$, and then $\dot{n}_j^r$ can be found from

$$\dot{n}_j^r\big|_{V_1} = x_j^r \dot{n}_r\big|_{V_1}. \quad (30)$$

Here, x_j^r is the mole fraction, which can be found using the mole numbers.

$$x_j\big|_{V_1} = \frac{n_j\big|_{V_1}}{\sum_j n_j\big|_{V_1}} \quad (31)$$

2.2.3 Heat Exchanger

The heat exchanger is considered as a standard counter-current heat exchanger with steady state heat transfer. The energy balance equation can be written as

$$\frac{dT_c}{dx} = \frac{UA}{\dot{m}_i \hat{c}_p^i L}\left(T_h - T_c\right), \quad (32)$$

and

$$\frac{dT_h}{dx} = \frac{UA}{\dot{m}_o \hat{c}_p^o L}\left(T_h - T_c\right) \quad (33)$$

where T_h, T_c are the temperatures of hot (outlet stream of the heat exchanger) and cold (inlet stream of the heat exchanger) streams at time t, respectively. U is the overall heat transfer coefficient of the heat exchanger and A is the total heat transfer area of the heat exchanger and $\hat{c}_p^i$ and $\hat{c}_p^o$ are the specific heat capacities of the inlet and outlet gas mixtures, respectively. L is the length of the heat exchanger and x is the position along the heat exchanger where $x = [0, L]$.

Assuming that $\dot{m}_o \hat{c}_p^o$ and $\dot{m}_i \hat{c}_p^i$ have the same values, and $\frac{UA}{\dot{m}_i \hat{c}_p^i}$ is independent of x, Eqs. 32 and 33 can be simplified further to give an explicit expression for the reactor inlet temperature as

$$T_r^i = \frac{T_i + \frac{UA}{\dot{m}_i \hat{c}_p^i} T_r^o}{1 + \frac{UA}{\dot{m}_i \hat{c}_p^i}}. \quad (34)$$

Similarly, the expression for the outlet temperature of the heat exchanger is

$$T_o = \frac{T_r^o + \frac{UA}{\dot{m}_i \hat{c}_p^i} T_i}{1 + \frac{UA}{\dot{m}_i \hat{c}_p^i}}. \quad (35)$$

3 Simulation Results and Discussion

3.1 Simulation Results

The mathematical model was simulated using the Python *odeint* solver with the use of the nominal values given in Appendix. Different number of volume compartments were tested to find the lowest number of volume compartments which sufficiently represents the system, and 150 volume compartments are selected. To obtain the oscillatory behavior of the temperature, the inlet temperature to the heat exchanger (T_i) was stepped down from 350°C to 230°C.

The temperature transient for 150 volume compartments is shown in Figure 3, depicting the change of temperature as uniform oscillations at the exit of the reactor with time, when the feed temperature (the inlet to the heat exchanger) was stepped down by 120°C. Initially, the reactor operated at steady state with a temperature of 350°C. Then at $t = 0.125$ hr, the temperature is reduced by 120°C. The system became unstable and showed oscillatory behavior. Temperature oscillations have a period of about 12 minutes and a maximum amplitude of about 320°C.

When the feed temperature decreases, the temperature at the reactor inlet also decreases due to the decreased heat transfer. This will affect the temperature at the exit of the reactor due to two mechanisms, which are the direct heat transfer from the gas and the change of heat of reaction of the exothermic reaction. The latter is known to be faster than the former (Morud and Skogestad, 1993). Therefore at first, the rate of ammonia conversion decreases leading to an increase of reactant concentration and total number of molecules in the first few reactor compartments, which will decrease the outlet temperature of each reactor compartment. This can be seen from the number of moles and the outlet temperature transient of the volume compartments 1 and 5 in Figure 4. A sudden reduction of number of molecules with the temperature reduction can be observed for the volume compartments 10 and higher. This may be due to the combined effect of the faster reduction of NH_3 molecules and slower increase of N_2 and H_2

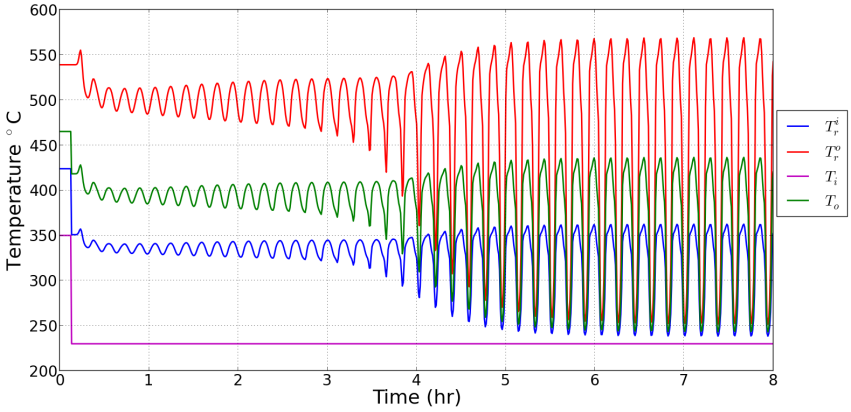


Figure 3. The temperature transient for 150 volume compartments, when a decrease in heat exchanger inlet temperature (T_i) from 350°C to 230°C was done at $t = 0.125$ hr.

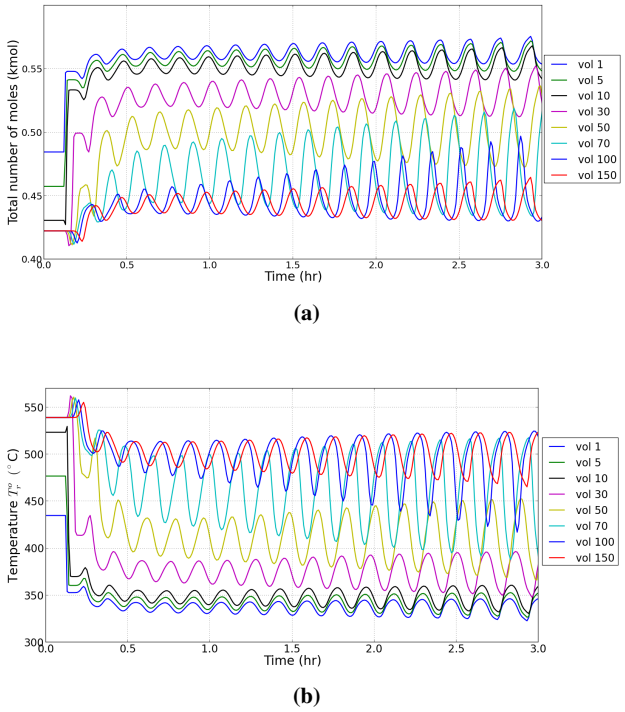


Figure 4. The total number of moles (4a) and the temperature (4b) with time, for different volume compartments along the reactor. Here 'vol' stands for the volume compartment number.

molecules. However, this gives a sudden increase to the exit temperature. The temperature of the reactor compartments along the reactor will increase due to the exothermic reaction of ammonia conversion. Therefore, the inlet temperature to the reactor will again increase by the heat transfer from the reactor exit streams. This dual effect of rate of reaction and the heat transfer will eventually result in an oscillatory behavior of number of moles in the reactor compartments leading to the same cyclic behavior in the outlet temperature of reactor compartments.

However, to obtain an optimum stabilized reactor performance, controlling of the temperature to the heat exchanger inlet will not be enough. The composition of feed gases, feed flow rate, feed temperature to the reactor inlet and the pressure along the reactor would be useful as monitoring measurements (Shah, 1967).

3.2 Comparison with Previous Work

The model developed by Naess et al. (1993) includes a reactor with three beds, an internal heat exchanger, an external heat exchanger, a compressor and a separator. A pressure drop is considered as a pressure drop across valves. Spatial discretization of states along the reactor beds is also done. Their model was verified using the plant data. However, the main objective was to test different control strategies.

The model used by Morud and Skogestad (1998) also consist of three beds in series with fresh feed make–up between each bed and pre–heating of feed with the effluent. Partial differential equations are used considering spatial discretization of temperature and the ammonia concentration in one direction. A dispersion coefficient is used for finding the finite heat transfer rate between the gas and the catalyst. The pre–heater is same as in this work, a steady state counter current heat exchanger, but the model used the *Number of Transfer Units* (NTU) approach with pre–heater efficiency. The temperature instability is obtained by changing the pressure of the reactor from 200 bar to less than 170 bar while the feed temperature was kept constant at 250°C. It is stated that the same behavior could be observed by changing the temperature from 250°C to about 235°C while keeping the pressure constant at 200 bar, which is also observed in this work.

The reactor system used by Rabchuk (2014) and Rabchuk et al. (2014) consists of a reactor and a heat exchanger as in this study. This is due to the assumption that the temperature oscillations occur due to the reactor–heat exchanger system, which is proven true. The mole numbers of species in the reactor and the heat flow through the heat exchanger is kept as states unlike in previous models,

where the concentration and temperature along the reactor beds were the states. The heat exchanger model includes dynamics and the *Logarithmic Mean Temperature Difference* (LMTD) approach with an approximation to the temperature difference. Instead of discretized partial differential equations, sets of ordinary differential equations have been used for 200 elementary volumes. The details of the reaction rate is not stated. Similar oscillatory behavior of temperature has been obtained by changing the temperature of the inlet to the system from 250°C to 200°C.

The topology used in this work is similar to that of Rabchuk et al. (2014), and simpler than the topology of most other work. Only the number of moles in the reactor compartments are kept as states via species balances. The heat exchanger model is explicit with respect to the temperature, which simplifies the model compared to other work where an implicit model based on LMTD is used. Assuming ideal gas, and perfectly controlled pressure allows for eliminating the energy balance to compute the exit flow rates. All the data with values and units are well–documented.

4 Conclusions

A mathematical model is developed for observing the dynamic behavior of an industrial ammonia synthesis reactor system which includes one reactor and a heat exchanger. All the data used in the simulation are taken from open literature and are presented in this work. The model is simple, but complete enough to satisfactorily reproduce the oscillatory behavior of the temperature of the reactor.

To obtain more accurate results, the model could be modified using the temperature dependent variables which are assumed as independent in this work and using more accurate catalyst activity values for the appropriate particle size of the catalyst.

Acknowledgment

Asanthi Jinasena thanks Anushka Perera for the help in learning Python.

References

R. F. Baddour, P. L. T. Brian, B. A. Logeais, and J. P. Eymery. Steady-state simulation of an ammonia synthesis converter. *Chemical Engineering Science*, 20(4):281–292, 1965. doi:10.1016/0009-2509(65)85017-5.

G. F. Froment, K. B. Bischoff, and J. D. Wilde. *Chemical Reactor Analysis and Design*. John Wiley & Sons, 2nd edition, 2010.

L. J. Gillespie and J. A. Beattie. The Thermodynamic Treatment of Chemical Equilibria in Systems Composed of Real Gases, I. An Approximate Equation for the Mass Action Function Applied to the Existing Data on the Haber Equilibrium. *Physical Review*, 36:743–753, 1930.

J. Morud. *Studies on the dynamics and operation of integrated plants*. PhD thesis, University of Trondheim, 1995.

J. Morud and S. Skogestad. The Dynamics of Chemical Reactors with Heat Integration. In *AIChE Annual Meeting*, pages 1–15, St. Louis, 1993.

J. C. Morud and S. Skogestad. Analysis of instability in an industrial ammonia reactor. *AIChE Journal*, 44(4):888–895, 1998. doi:10.1002/aic.690440414.

A. Murase, H. L. Roberts, and A. O. Converse. Optimal Thermal Design of an Autothermal Ammonia Synthesis Reactor. *Industrial & Engineering Chemistry Process Design and Development*, 9(4):503–513, 1970. doi:10.1021/i260036a003.

L. Naess, A. Mjaavatten, and J.-O. Li. Using dynamic process simulation from conception to normal operation of process plants. *Computers & Chemical Engineering*, 17(5-6):585–600, 1993. doi:10.1016/0098-1354(93)80046-P.

K. Rabchuk. *Stability map for ammonia synthesis reactors*. Master's thesis, Faculty of Technology, Telemark University College, Norway, 2014.

K. Rabchuk, B. Lie, A. Mjaavatten, and V. Siepmann. Stability map for Ammonia Synthesis Reactors. In *55th Conference on Simulation and Modeling (SIMS 55)*, pages 159–166, 2014.

M. Shah. Control Simulation in Ammonia Production. *Industrial & Engineering Chemistry*, 59(1):72–83, 1967. doi:10.1021/ie50685a010.

C. P. P. Singh and D. N. Saraf. Simulation of ammonia synthesis reactors. *Industrial & Engineering Chemistry Process Design and Development*, 18(3):364–370, 1979. doi:10.1021/i260071a002.

V. B. Singh. *Modelling of Quench Type Ammonia Synthesis Converters*. Master's thesis, Department of Chemical Engineering, Indian Institute of Technology Kanpur, India, 1975.

C. van Heerden. Autothermic Processes, Properties and Reactor Design. *Industrial & Engineering Chemistry*, 45(6):1242–1247, 1953. doi:10.1021/ie50522a030.

Appendix: Data

Parameters and operating conditions used for the simulation:

Parameters

A	Heat transfer area (Morud, 1995)	283 m^2
$\tilde{c}_p$	Molar heat capacity of gas mixture (Morud, 1995)	35500 $\frac{J}{kmol \cdot K}$
$\hat{c}_{p,c}$	Specific heat capacity of catalyst (Morud, 1995)	1100 $\frac{J}{kg \cdot K}$
$C_{p,c}$	Total heat capacity of catalyst $m_c \hat{c}_{p,c}$	$138.4 \times 10^6 \frac{J}{K}$
ΔH_r	Enthalpy of the reaction (Rabchuk et al., 2014)	$-92.4 \times 10^6 \frac{J}{kmol}$
E_-	Activation energy of reverse reaction (Murase et al., 1970; Morud and Skogestad, 1998)	$1.98464 \times 10^8 \frac{J}{kmol}$
ε	Void fraction of catalyst (Rabchuk et al., 2014)	0.42
f	Catalyst activity factor (Morud and Skogestad, 1998)	4.75
k_-^0	Pre–exponential factor of reverse reaction (Murase et al., 1970; Morud and Skogestad, 1998)	$2.5714 \times 10^{16} \frac{kmol \cdot atm^{\frac{1}{2}}}{m^3 \cdot h}$
m_c	Total mass of catalyst $\rho_c V$	125840 kg
M_{Ar}	Molar mass of Ar atom	39.95 $\frac{kg}{kmol}$
M_{H_2}	Molar mass of H_2 molecule	2.016 $\frac{kg}{kmol}$
M_{N_2}	Molar mass of N_2 molecule	28.02 $\frac{kg}{kmol}$
M_{NH_3}	Molar mass of NH_3 molecule	17.034 $\frac{kg}{kmol}$
N	Number of reactor compartments (Decided after a few trials)	150
ν	Stoichiometric matrix [H_2 N_2 NH_3 Ar]	[-3 -1 2 0]
p^σ	Atmospheric pressure	1.01325×10^5 Pa
R	Universal gas constant	8314 $\frac{J}{kmol \cdot K}$
ρ_c	Packing density of catalyst (Rabchuk et al., 2014)	2200 $\frac{kg}{m^3}$
U	Overall heat transfer coefficient (Morud, 1995)	$1.9296 \times 10^6 \frac{J}{h \cdot m^2 \cdot K}$
V	Volume of the reactor (Rabchuk et al., 2014)	57.2 m^3

Operating Conditions

$\dot{m}_i$	Mass flow rate - reactor inlet (Rabchuk et al., 2014)	67.6 $\frac{kg}{s}$
p	Controlled reactor pressure (Rabchuk et al., 2014)	178×10^5 Pa
T_i	Feed temperature (heat exchanger inlet) (Rabchuk et al., 2014)	350 °C
$x^i_{H_2}$	Mole fraction of H_2 at reactor inlet (Rabchuk et al., 2014)	0.6972
$x^i_{N_2}$	Mole fraction of N_2 at reactor inlet (Rabchuk et al., 2014)	0.24
$x^i_{NH_3}$	Mole fraction of NH_3 at reactor inlet (Rabchuk et al., 2014)	0.0212

Comparison of OpenFOAM and ANSYS Fluent

Prasanna Welahettige, Knut Vaagsaether

Department of Process, Energy and Environmental Technology
University College of Southeast Norway
Porsgrunn, Norway
prasanna.welahetti@gmail.com

Abstract

Gas-gas single phase mixing were numerically evaluated with static mixer and without static mixer using OpenFOAM and ANSYS Fluent codes. The main goal was the gas-gas mixing simulation comparison between ANSYS Fluent and OpenFOAM. The same ANSYS mesh was used for each case in both codes. The "reactingFoam" solver and species transport models were used for handling the species in OpenFOAM and ANSYS Fluent respectively. The reactingFoam solver is a transient solver and ANSYS Fluent was simulated at steady state condition. Standard k-ε model was used to predict the turbulence effect in both computational fluid dynamics codes. OpenFOAM gave a higher mixing level compared to ANSYS Fluent. Chemical species momentum predictions are more diffusion in OpenFOAM and more convective in ANSYS Fluent.

Keywords: OpenFOAM, ANSYS Fluent, CFD, mass fraction, standard deviation, mixing

1 Introduction

OpenFOAM (OF) is a free computational fluid dynamics (CFD) code developed by OpenCFD Ltd and OpenFOAM Foundation. This CFD code is getting well known in academic and industrial sector due to a broad range of fluid dynamics applications, open source, no limitation for parallel computing, and no limitation of number of species in the chemistry models (Lysenko et al., 2013). ANSYS Fluent (AF) is a commercial code and it is developed for the CFD simulation with powerful graphical user interface. It was developed and maintained by ANSYS Inc. It is possible to achieve the same result from both OF and AF by examining "the exterior flow field around simplified passenger sedan" geometry (Ambrosino and Funel, 2006). Reynolds-averaged Navier-Stokes equations (RANS equations) with the k-ε model can be used for both OF and AF CFD codes. OF predicts more accurate results for the velocity and AF predicts more accurate results for the turbulent kinetic energy (Balogh et al., 2012). According to the analyze of "turbulence separated flows" using OF and AF, turbulence models give closely equal results from the both CFD codes (Lysenko et al., 2013). The main objective of this work is to compare the OF simulation ability with the AF simulation for gas-gas single phase mixing. OpenFOAM 2.4.0 and ANSYS Fluent R16.2 academic version were used for the simulations. These geometry designs are unique models for an industrial application and there are no published experimental results.

2 Solvers selection

The solver selection is an important step in this study. Chemical and physical properties of the system are considered for the solver selection. The continuity equation is required to keep the mass balance constantly. The momentum equation contributes to calculate the velocities and the pressures. This system was operated at an isothermal condition. However, the energy equation was required to predict the densities at the operating pressure and temperature. A multispecies model was required because of air-ammonia mixing. The flow was turbulent. Therefore, a turbulence model was required to predict the turbulence properties. The fluid flow highly interacts with walls and mixer plates. Therefore, a wall treatment method is required.

2.1 The species transport model

A species transport equation is given in Equation-1. It describes the convection and the diffusion of the species i for a unsteady condition without a reaction (Fluent, 2006; Versteeg and Malalasekera, 2007).

$$\frac{\partial \rho Y_i}{\partial t} + \mathrm{div}(\rho U Y_i) = -\mathrm{div}(J_i) + S_i \quad (1)$$

Here, ρ is the density of species i, Y_i is the mass fraction of species i, U is the three dimensional velocity components, t is time, J_i is the diffusion flux of species i, and S_i is the source term of species i. The mass diffusion in a turbulent flow is given as,

$$J_i = -\rho D_i \mathrm{grad}(Y_i) - \frac{\mu_t}{Sc_t}\mathrm{grad}(Y_i) \quad (2)$$

Here μ_t is the turbulent viscosity and Sc_t is the turbulent Schimidt number. The reaction term was neglected because of only the air-ammonia mixing.

2.2 Solvers parameters comparison

A comparison of solvers parameters is shown in Table 1. OF was a transient simulation and AF simulation was a steady state simulation. Most of the OF chemical solvers are transient solvers. However, AF has both transient and steady state solvers for the chemical species. A steady state solver was selected for AF to save the computational time. Ammonia and air are in the AF chemical species database. However, OF does not have a database for chemical species. Therefore, molecular weight, heat capacity, heat of fusion, dynamic

viscosity and Prandtl number were added as the species properties.

"reactingFoam" is a compressibility based solver. Compressibility is defined as inverse of the multiplication of the temperature and the universal gas constant $(RT)^{-1}$ (Greenshields, 2015). The temperature was a constant and the fluid mixer gases were considered as perfect gases in this study. Therefore, the compressibility was constant for the "reactingFoam" simulations. The pressure velocity coupling was handled in two different ways in both CFD codes. PIMPLE algorithm was created by merging the Pressure Implicit with Splitting of Operators (PISO) algorithm and the Semi Implicit Method for Pressure Linked Equations (SIMPLE) algorithm (Versteeg and Malalasekera, 2007). The PIMPLE algorithm operates at PISO mode if the non-orthogonal correction number higher than one (Greenshields, 2015). In this simulation, the non-orthogonal correction number was equal to two. Therefore the pressure velocity coupling was handled by the PISO algorithm in OF. The PISO contains an extra correction step compare to the SIMPLE. Therefore, OF calculations were with an extra correcting than AF calculations in these simulations. These algorithms should not influence the solutions. It influences to the solution calculation methods.

Table 1. Simulation settings comparison between OF and AF.

	OF	AF
Solver type	Compressibility based Transient	Pressure based Steady state
Models	Energy equation Viscous – standard k-ε reactingFoam (solver)	Energy equation Viscous – standard k-ε Species transport
Materials	NH_3, Air	NH_3, Air
Solution Method	PIMPLE limitedLinear	SIMPLE Second order upwind
Solution control	Selected time step	Default under relaxation factors

3 Methods

The basic sketch is shown in Figure 1 to demonstrate the flow directions. In this sketch, the ammonia-injecting pipe and the static mixer modules are not presented.

3.1 Geometries drawing

The ANSYS DesignModeler (DM) tool was used to make the drawings. The purpose of the designs was to mix air and ammonia in the horizontal circular tube. The geometries were drawn based on the data given by YARA International ASA. The parts of the geometry are ammonia injecting pipe, static mixer plates and cylindrical tube walls.

The geometries were drawn in three cases as "without ammonia injecting pipe and without static mixture modules" (case-1), "with ammonia injecting pipe and without static mixer modules" (case-2) and "with ammonia injecting pipe and with static mixer modules" (case-3).

The total length of the large pipe is 5.5 m and diameter is 0.996 m. The length of the ammonia-injecting pipe is 1.473 m and diameter is 0.25 m (the small cylindrical pipe with holes). Figure 2 shows the geometries for above-mentioned three cases. There are a 10 numbers of holes in the ammonia-injecting pipe and the diameter of a hole was 80 mm. The mixer plates were drawn as zero thickness walls to reduce the number of elements in the meshes. There were four number of mixer modules and the mixer plates arrangements were similar in each module but module 2 and 4 were rotated around the pipe central axis by 90°.

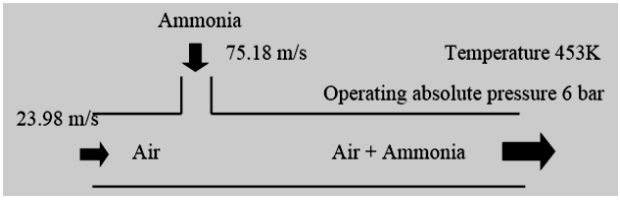


Figure 1. Air and ammonia inlets, outlet and flow directions.

3.2 Mesh generation

The same mesh was used for simulation in both CFD codes in each case for an accurate comparison. Three meshes were generated with respect to the above mentioned three geometry cases. These meshes are shown in Figure 3 and the element types are shown in Table 2. The minimum and maximum element sizes are 2.88 mm and 288 mm respectively for all 3 cases. More tetrahedral elements are added when the complexity is increased in the geometries. Inflation layers are added to better prediction in near the walls and the mixer plates. The total number of cells is increased by five times due to inclusion of the static mixer.

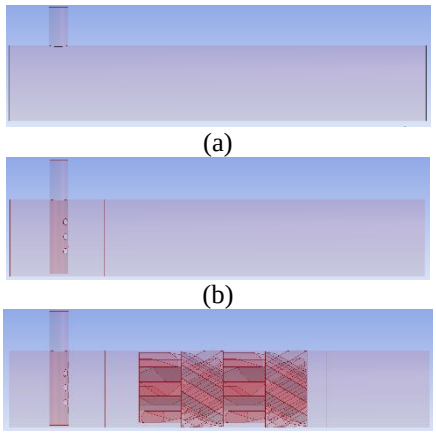
3.3 Boundary conditions and initial values for OF and AF

The boundary conditions were named as "velocity_inlet_air", "velocity_inlet_nh3", "pressure_outlet" and "walls". The velocity inlet type was selected by assuming an incompressible fluid. A comparison of boundary conditions is shown Table 3. The air inlet velocity was 23.98 m/s. The turbulence intensities were 1.7 % and 1.9 % in air inlet and ammonia inlet respectively. All the outer cylinder surfaces and mixer plates were considered as the walls. The walls were assumed as "no slip" condition and there was no heat transfer through the walls.

3.4 Mixing Evaluation Method

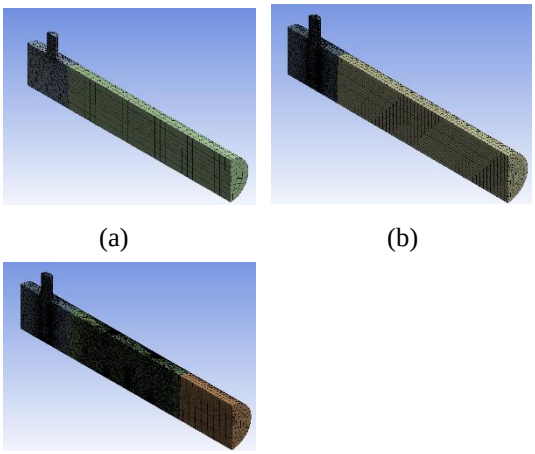
Mixing evaluation can be done using a "mixer parameter" (Kok and van der Wal, 1996). It is based on the standard deviation of species mass fraction. The sample points are considered in a line. This model was

considered as the basic model for the mixing evaluation in this study. The same basic mixing evaluation method was applied for this study in a different manner to evaluate the mixing of ammonia and air. Ten number of mixing evaluation planes were defined after the static mixer modules. The mixing evaluation planes were circular cross sections of the large cylindrical pipe as shown in Figure 4. The gap between two subsequent planes is 100 mm. These mixing evaluation planes are located after the static mixer modules to evaluate the fluid coming out from static mixer modules. Standard deviation values of ammonia mass fraction were calculated for the each mixing evaluation planes. All cells available in a mixing evaluation plane were considered to calculate a standard deviation value. Ten number of standard deviation values were calculated for a geometry model. Standard deviation values were plotted against x/D values for the each geometry models.



(a) (b) (c)

Figure 2. Geometry; (a) Case-1, (b) Case-2, (c) Case-3.



(a) (b) (c)

Figure 3. Mesh cross sectional view; (a) Case-1, (b) Case-2, (c) Case-3.

Here, x is the axial position from air inlet and D is the large circular pipe diameter. If the curve gives a lower values line, the design gives a better mixing than the others. If the curve gives a higher values line, the design gives a lower mixing than the others. The mixing evaluation model proposed in this study was more accurate than the basic model and easy to compare with the similar geometry models.

Table 2. Elements details of meshes.

Type of cells	Case-1	Case-2	Case-3
Hexahedra	1410	10192	4505
Prisms	60	98	136
Pyramids	47	208	265
Tetrahedral	8321	235100	1214040
Total cells	9838	245598	1218946

This method gives an overall idea about the mixing. All the cells available in each mixing evaluation planes contribute for the calculations. This method is more accurate if the gap between two planes is reduced and the numbers of evaluation planes are increased.

Table 3. Boundaries comparison between OF and AF.

Boundary		**OF**	**AF**
Name	***Variable***		
Velocity_inlet_nh3	Velocity	fixedValue - 75.18 m/s	75.18 m/s
	Pressure	zeroGradient	Gauge Pressure–0
	NH_3	fixedValue – 1	1
Pressure_outlet	Velocity	PressureInletOutletVelocity, $internalField	-
	Pressure	TotalPressure	Gauge Pressure–0

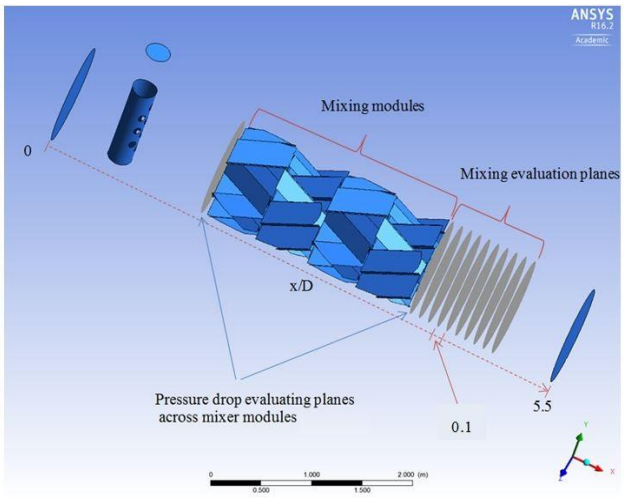


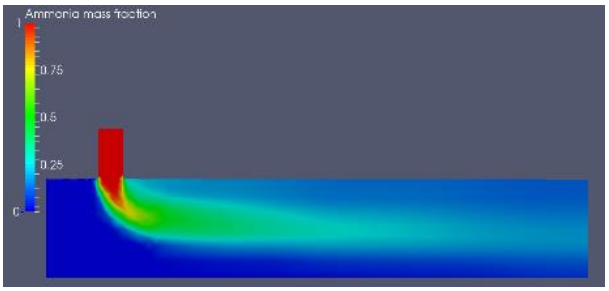
Figure 4. Mixing evaluation planes.

4 Results

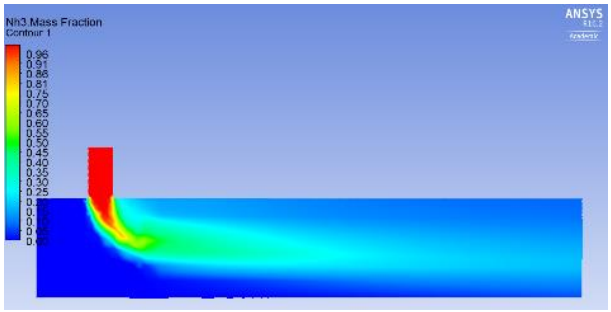
Observations of the mixing behavior were the main objective in results analysis. Computational time optimization was not focused in this study.

4.1 Case-1

Ammonia mass fraction contours are shown in Figure 5 for both CFD codes. The OF simulation showed a bit more diffusive flow patterns compared to the AF. The AF simulation showed longer convective ammonia channels. These channels became less diffusive toward the outlet in AF compared to OF. A higher mixing level was shown in the OF simulation compared to AF as shown in Figure 6 (OF showed lower standard deviation values). It was due to a higher diffusive behavior of OF simulation. Mixing level was further increased towards the outlet because the residence time increased. This was basically due to macro mixing. When the residence time increases, the interaction between molecules further increases.



(a)



(b)

Figure 5. Ammonia mass fraction contours case -1; (a) OF, (b) AF.

4.2 Case-2

Theoretically, a perfect mixing gives a stoichiometric mixture. The stoichiometric ammonia mass fraction was equal to 0.16 for these simulations. Ammonia mass fraction was plotted along the central axis line in large circular pipe as shown in Figure 7.

Ammonia mass fraction was zero in $-0.65 < x/D < 0$ range, because of large upstream airflow. This implies that there was no backflow of ammonia. At the ammonia inlet tube $-0.125 < x/D < 0.125$, ammonia mass fraction was equal to one. This means that there was no initial air remaining in the ammonia-injecting pipe at the steady state. There was an increment in ammonia mass fraction close to x/D = 0.35. This was due to the position of monitoring line, which was placed between two conservative holes of the injector. At x/D > 2.5, ammonia mass fraction was a constant value in the AF simulation. However, there was a fluctuation in the OF simulation. This was basically due to "wiggles" formation in OF simulation. These "wiggles" patterns generated with transient simulation in OF. As well as, the "wiggles" helped to increase the mixing in OF compared to AF. However, the OF result was fluctuating around the AF results. Therefore the average ammonia mass fraction value of OF was almost similar to the AF result.

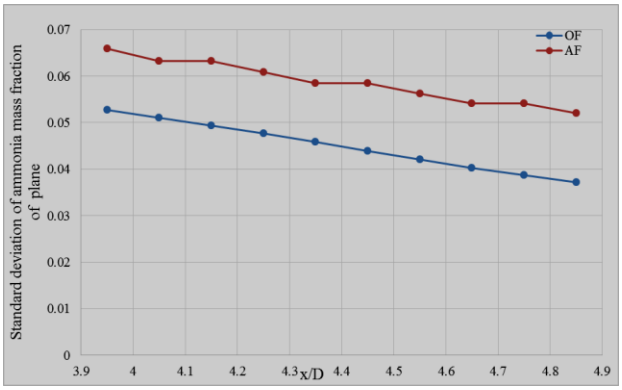


Figure 6. Mixing evaluation – case-1 (standard deviation of ammonia mass fraction).

4.3 Case-3

Ammonia mass fraction contours are shown in Figure 8. Both CFD codes show similar contours in this case. Standard deviation of ammonia mass fraction is shown in Figure 9. Case-3 showed lowest standard deviation values from both CFD codes compared to the previous cases (case-1 and case-2). This implies that the static mixer has improved the mixing. OF simulation result showed higher mixing (lower standard deviation values) in case-3 compare to AF.

Ammonia mass fraction contours at x/D = 4.85 (this location is the 10th mixing evaluation plane in Figure 4) is shown in Figure 10. The range of ammonia mass fraction was shown from 0.1 to 0.2 to compare with perfect mixing. A perfect mixing was given at ammonia mass fraction equal to 0.16 from a stoichiometric mixture. Higher perfect mixing regions were shown in OF compared to AF. Further, ammonia mass fraction (0 to 1 range) was plotted along the vertical line (the vertical line is shown in Figure 10) at x/D = 4.85 as shown in Figure 11. Here y is the vertical axis position (y negative means bellow the center of the circular plane). It also showed that OF simulation showed higher ammonia mass fraction regions than AF.

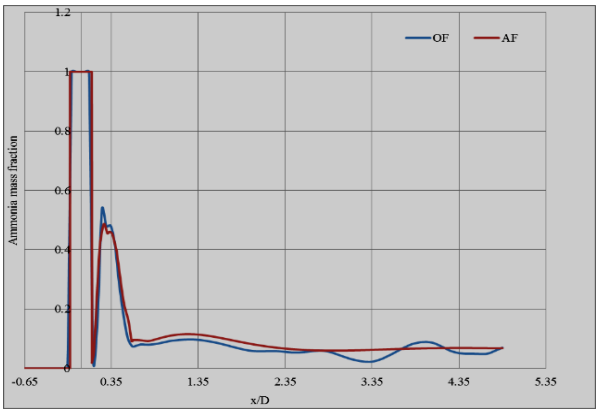
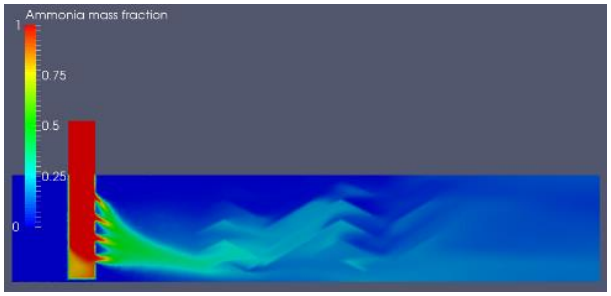
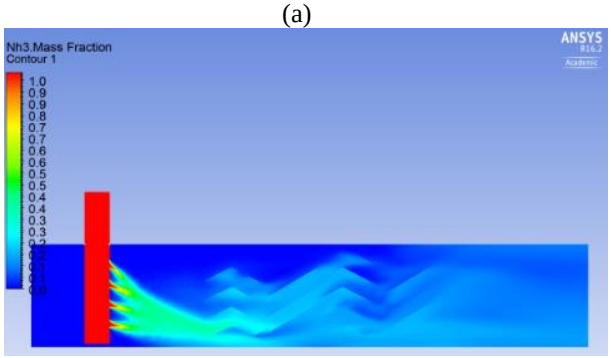


Figure 7. Ammonia mass fraction along the central axis – case-2.



(a)



(b)

Figure 8. Ammonia mass fraction case-3; (a) OF, (b) AF.

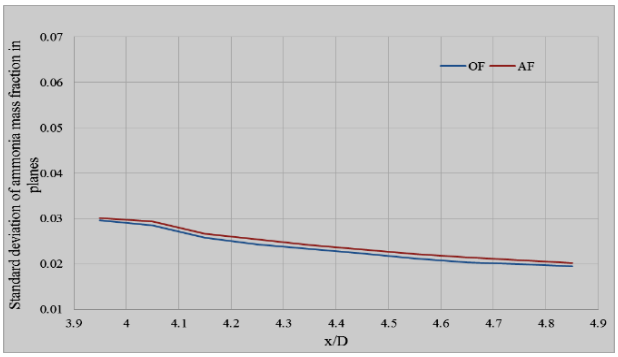
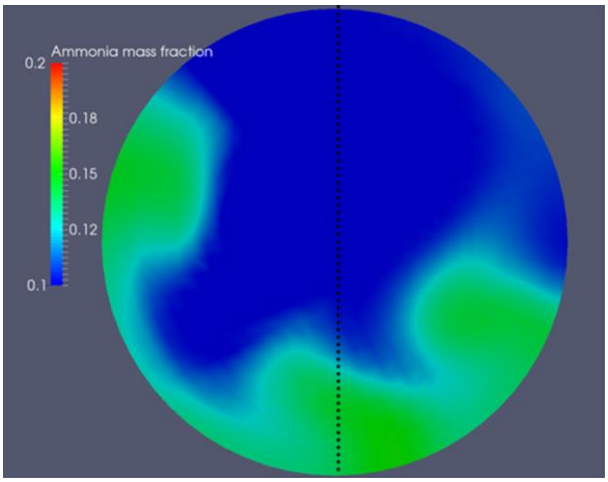
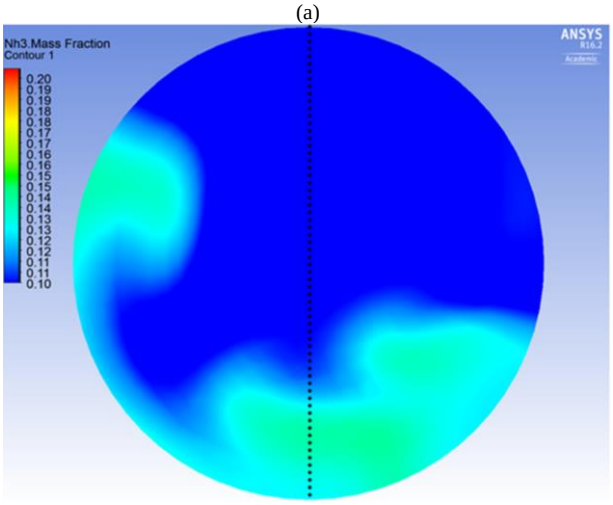


Figure 9. Mixing evaluation - case 3 (standard deviation of ammonia mass fraction).

The average turbulent kinetic energy of the mixing evaluation planes (the mixing evaluation planes are shown in Figure 4) are shown in Figure 12. Higher average turbulent kinetic energy was predicted by OF than AF. Turbulent kinetic energy is reduced toward the outlet due to decay of turbulence.



(a)



(b)

Figure 10. Ammonia mass fraction contours at x/D = 4.85; (a) OF, (b) AF.

5 Discussion

5.1 Mesh quality requirement comparison between OF and AF

Quality of the mesh is one of the key parameter to control the accuracy and the stability of the computational schemes (Fluent, 2006). Skewness is used to check the quality of the mesh in generally. Case-3 geometry was used to simulate the different skewness meshes in this experiment. OF was able to simulate only less than 0.78 maximum skewness meshes. However, AF simulated maximum skewness up to 0.93. This means that AF has more ability to handle low quality mesh than OF.

5.2 Jet mixing comparison between OF and AF

When a fluid mixes with another fluid, velocity shear layers are created between two fluids (Yuan et al., 2004). Ammonia coming out from the small holes can be considered as non-premixed turbulent jet flows. Ammonia mass fraction contours at the first hole is shown in Figure 13 for the both CFD codes. Concentration of ammonia was decreased, when mixing occurred. AF showed comparatively longer shear boundaries. OF shear boundaries were wider than AF boundaries. Figure 14 shows ammonia mass fraction across the jet in a vertical line. The vertical lines are shown in Figure 13. These lines were selected at same distance from the injector pipe. Ammonia mass fraction around the jet was larger in OF while the ammonia mass fraction at the center of the jet was larger in AF. This implies that, higher diffusion was shown in OF and higher convection was shown in AF. This was a reason for the OF showed higher mixing because of higher diffusion effect.

5.3 Vortex street formation comparison

Flow around the cylinder creates turbulent vortexes at Re > 3.5 $\times 10^6$ range (Fluent, 2006). The average Reynolds number was 3.8 $\times 10^6$ in these simulations. Vortex formation comparison around ammonia injecting pipe is shown in Figure 15 (velocity contours). OF showed a higher vortex street formation compared to AF. These high number of vortex streets formation caused the higher mixing level in OF compared to AF. Vortex streets are averaging in steady state simulation but not in transient simulations. OF simulation was in transient mode and AF simulation was in steady state mode. Therefore, transient simulation increased the mixing compared to the steady state simulation.

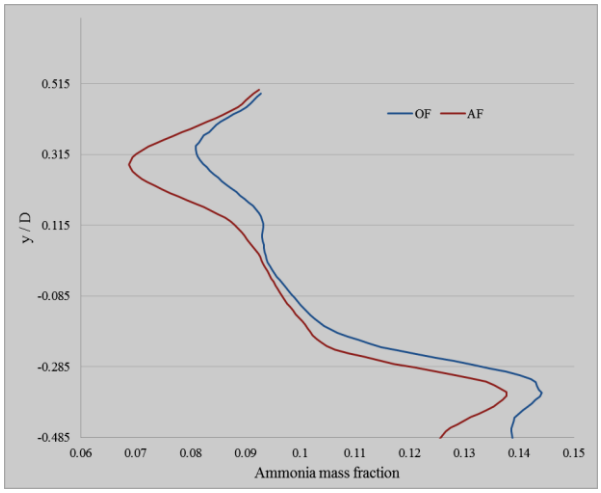


Figure 11. Ammonia mass fraction along the vertical lines at x/D = 4.85(y negative is below the center of plane and vertical lines are shown in Figure 10).

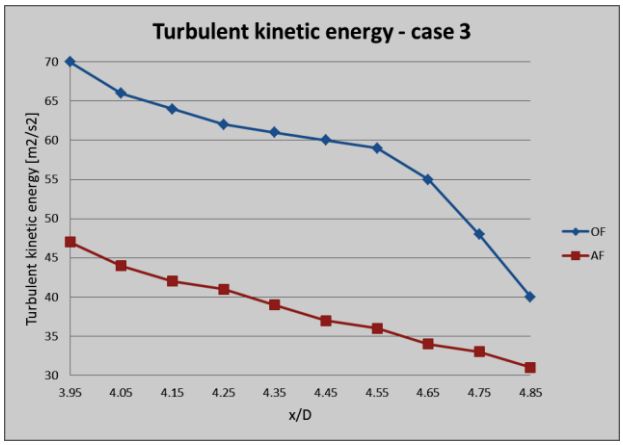
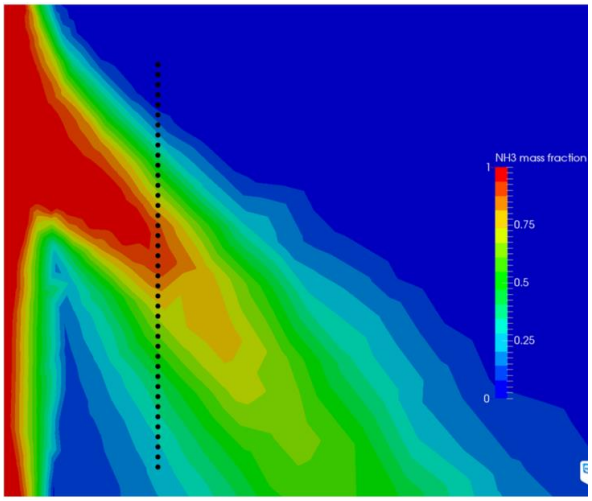
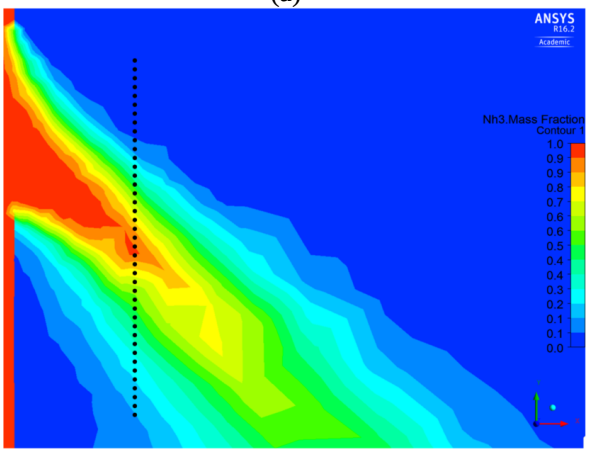


Figure 12. Turbulent kinetic energy comparison – case 3 (Average turbulent kinetic energy in mixing evaluation planes).



(a)



(b)

Figure 13. Jet mixing at first hole in injector; (a) OF, (b) AF.

5.4 Eddy prediction comparison between OF and AF

The standard k-ε model was used by both CFD codes for turbulence prediction. Figure 16 shows large and small eddies in OF and AF. The same locations were

considered for the comparison in both codes. The large eddies were possible to see in between the cylindrical wall and the mixer plates. More large eddies were predicted by AF than OF. However, small eddies were comparatively equally predicted in the both codes.

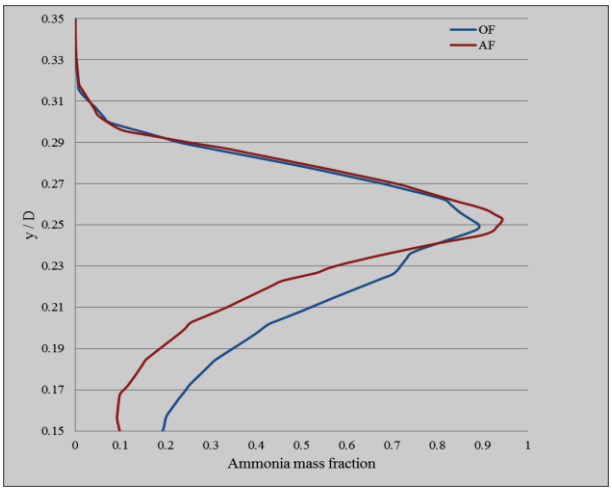
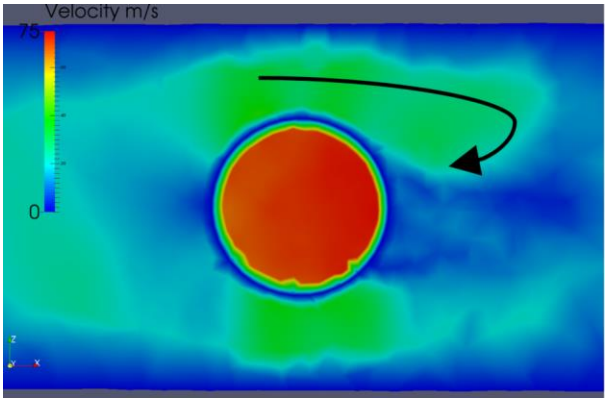
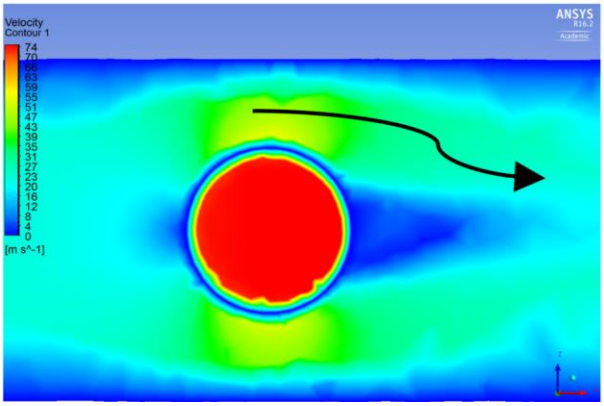


Figure 14. Ammonia mass fraction across a jet (mass fractions were plotted along the vertical lines those lines are shown in Figure 13).

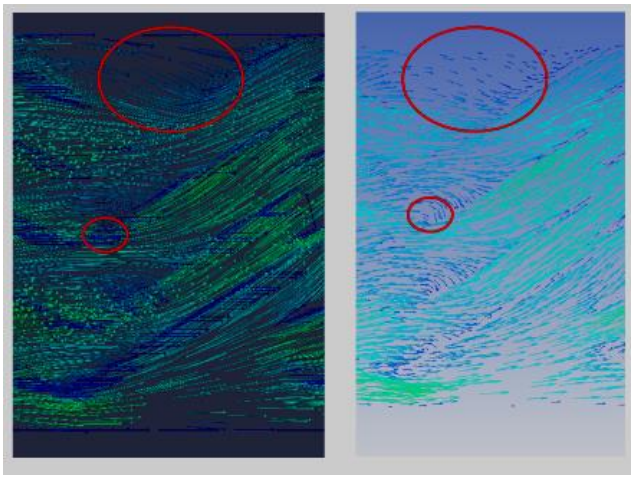


(a)



(b)

Figure 15. Vortex street comparison – Velocity contours around ammonia injecting pipe; (a) OF, (b) AF.



(a) (b)

Figure 16. Large and small eddies (large circles and small circles show large eddies and small eddies respectively); (a) OF, (b) AF.

5.5 Turbulence kinetic energy effect comparison between OF and AF

Energy cascade path starts with main flow energy to large eddies, large eddies energy to small eddies, small eddies energy to smallest eddies and finally to internal energy (Versteeg and Malalasekera, 2007). OF simulation showed higher turbulent kinetic energy than AF in this study. These high turbulent kinetic energies helped to create more vortices in OF according to the energy cascade principle. Because of this, higher mixing was shown in OF compared to AF.

5.6 Summary of OF and AF results comparison

A summary of result comparison is shown in Table 4.

Table 3. Summary of OF and AF results comparison.

Static mixer model	Description
Without static mixer	OF gives higher mixing
With static mixer	Both codes give on average equal mixing

6 Conclusions

Both codes used same constant values for k-ε equation and same initial values. Higher turbulent kinetic energy was predicted from OF compared to AF. As well as higher diffusive properties was shown in OF compared to AF. Those two reasons were mainly involved to predict the higher mixing in OF compared to AF. OF simulation is required higher quality mesh compare to AF. As an example, 0.93 maximum skewness mesh gives a converged result in AF but same mesh gives a diverged result in OF. This implies that finer mesh is

required for OF. Both codes used the ANSYS meshes, which are designed for AF.

Acknowledgment

The authors express their thanks to Luigi Serraiocco, Jakub Bujalski and YARA International ASA for useful technical support.

References

F. Ambrosino and A. Funel. OpenFOAM and Fluent features in CFD simulations on CRESCO High Power Computing system. In *Final Workshop of Grid Projects, PON RICERCA*, 1 – 4, 2006.

M. Balogh, A. Parente, and C. Benocci. RANS simulation of ABL flow over complex terrains applying an Enhanced k-ε model and wall function formulation: Implementation and comparison for fluent and OpenFOAM. *Journal of Wind Engineering and Industrial Aerodynamics*, 104-106: 360–368, 2012. doi:10.1016/j.jweia.2012.02.023

ANSYS Fluent. *Fluent 6.3 Documentation*. Fluent Inc., Lebanon, NH, 2006.

C. J. Greenshields. *Openfoam user guide*. OpenFOAM Found. Ltd, version 3, 2015.

J. B. Kok and S. van der Wal. Mixing in T-junctions. *Applied Mathematical Modelling*, 20: 232–243, 1996. doi:10.1016/0307-904X(95)00151-9

D. A Lysenko, I. S. Ertesvåg, and K. E. Rian. Modeling of turbulent separated flows using OpenFOAM. *Computers & Fluids*, 80: 408–422, 2013. doi:10.1016/j.compfluid.2012.01.015

H. K. Versteeg and W. Malalasekera. *An introduction to computational fluid dynamics: the finite volume method*, 2nd ed. Pearson Education Ltd, 2007.

C. C. L. Yuan, M. Krstić, and T. R. Bewley. Active control of jet mixing. *IEE Proceedings-Control Theory Appl*, 151(6): 763–772, 2004. doi:10.1080/14685248.2014.997244

Impact of Particle Diameter, Particle Density and Degree of Filling on the Flow Behavior of Solid Particle Mixtures in a Rotating Drum

Sumudu Karunarathne[1] Chameera Jayarathna[2] Lars-Andre Tokheim[1]

[1]Department of Process, Energy and Environmental Technology, University College of Southeast Norway, Norway
{sumudu.karunarathne,lars.a.tokheim}@usn.no
[2]Tel-Tek, Norway, chameera.jayarathna@tel-tek.no

Abstract

Two-dimensional CFD simulations were performed to investigate the impact of density, particle diameter and degree of filling on the flow behavior of a solid particle mixture in a transverse plane of a rotary drum. The Eulerian approach with the kinetic theory of granular flow was used to simulate granular phases of $CaCO_3$ and Al_2O_3 under the rolling mode. The volume fractions of each phase reveal that, under the considered conditions, the particle size has a greater impact on segregation than the density. Larger particles are collected at the bottom of the rotating drum while smaller particles move more into the mid-section of the bed. The active layer is responsible for the segregation owing to the trajectory mechanism. Particle segregation due to percolation is more dominant than segregation due to condensation. In addition to that, solids volume fraction variations in the moving bed indicate that the influence of the degree of particle filling made no significant impact on the degree of mixing in the rotating drum.

Keywords: rotating drums, segregation, granular flow, rolling mode, active layer

1 Introduction

Granular particles are processed in many ways, and for various purposes, in the industry. Rotary drums are widely used in kilns, mixers, dryers and reactors (Ding et al., 2001). In pyro-processing of materials, rotary kilns are used due to its effective mixing and heat transfer capabilities. Transverse and longitudinal mixing occur in rotating drums. In the transverse motion, six types of particle bed behavior are observed, depending on Froude number, filling degree, wall friction coefficient, ratio of particle to cylinder diameter, angle of internal friction and dynamic angles of repose (Yin et al., 2014). The bed motions are known as slipping, slumping, rolling, cascading, cataracting and centrifuging, and the rolling mode is preferred in both horizontal drums and kiln operation due to efficient mixing performance (Demagh et al., 2012; Boateng et al., 2008).

Homogeneous mixtures are required in most industries to maintain the quality of the end product. However, during processing particles may segregate when they are subjected to motion. Segregation in process equipment can be seen in many industrial applications owing to variations in the characteristics of the particles that are processed. Different particle size, shape, density, roughness and resilience are the main causes of the segregation (Boateng and Barr, 1996; Henein et al., 1983). In rotating drums, segregation can be observed in both radial and longitudinal directions. Trajectory segregation is possible in the flow regimes of slumping, rolling and cataracting, in which finer particles will be concentrated in the mid-chord section (Boateng and Barr, 1996). Percolation and condensation are two other segregation mechanisms that may occur in rotating drums. Percolation is caused by differences in particle size, whereas condensation is due to variation in particle densities. One of these mechanisms may dominate in the active layer of the moving bed, leading to segregation of particles during drum rotation (Chen et al., 2016).

Several experimental and numerical simulations have been done to investigate particle dynamics in rotating drums. Boateng and Barr (1996) developed a mathematical model to investigate mixing and segregation in a transverse plane of a rotary kiln. Their model was capable of predicting the active layer depth and the velocity distribution. The discrete element method (DEM) is a promising technique to investigate particle motion in many industrial applications. Longitudinal and transverse mixing in rotary kilns were studied by Finnie et al. (2005) using the DEM approach. Mixing in the longitudinal direction was explained by a one dimensional diffusion equation, and transverse mixing was determined by using an “entropy” like quantity proposed by Schutyser et al. (2001). Chen et al. (2016) performed a numerical simulation to investigate radial mixing and segregation of a granular bed considering particle size and density variations. It was concluded that a homogeneous mixture can be obtained by achieving an equilibrium state between percolation and condensation by controlling the volume ratio or the density ratio of the granular particles.

This study focuses on two-dimensional (2D) numerical simulations of mixing and segregation of two granular phases in a rotating drum. Two granular materials, calcium carbonate ($CaCO_3$) and aluminum oxide Al_2O_3, were considered in the simulations. The Eulerian approach along with the kinetic theory of granular flow were used to analyze the particle dynamics and segregation in the transverse plane. The focus was on understanding how density, particle size

and degree of filling affect the particle segregation in a rotating drum.

A two-dimensional geometry was created to mimic the characteristics of the transverse plane of a rotating drum. The numerical calculations were performed by the CFD software ANSYS FLUENT 16.2.

2 Model Description

The Eulerian approach is applicable to flows with N number of phases. Here, the governing equations for the Euler-Euler model and the kinetic theory of granular flow are discussed for a solid mixture.

2.1 Governing Equations in the Euler-Euler Method

2.1.1 Continuity Equations

Conservation of mass in the flow is represented by continuity equations for the gas phase and the solid phases,

$$\frac{\partial}{\partial t}\left(\varepsilon_g \rho_g\right)+\nabla\cdot\left(\varepsilon_g \rho_g v_g\right)=0 \tag{1}$$

$$\frac{\partial}{\partial t}\left(\varepsilon_S \rho_S\right)+\nabla\cdot\left(\varepsilon_S \rho_S v_S\right)=0 \tag{2}$$

Here, ρ is density, v is velocity, ε is volume fraction and t is time. S and g refer to the solids phase and the gas phase, respectively.

2.1.2 Momentum Equations

The influence of viscous, pressure and gravity forces on the dynamics of the gas and the solid particles is described by momentum equations. The momentum equations for the gas phase and the solid phases are written as (Demagh et al., 2012; Azadi, 2011):

$$\frac{\partial}{\partial t}\left(\varepsilon_g \rho_g v_g\right)+\nabla\cdot\left(\varepsilon_g \rho_g v_g v_g\right)=-\varepsilon_g \nabla P_g+\varepsilon_g \rho_g g - \sum_{s=1}^{N} k_{gs}\left(v_g-v_s\right)+\nabla\cdot\left(\varepsilon_g \tau_g\right) \tag{3}$$

$$\frac{\partial}{\partial t}\left(\varepsilon_s \rho_s v_s\right)+\nabla\cdot\left(\varepsilon_s \rho_s v_s v_s\right)=-\varepsilon_s \nabla P_g+\varepsilon_s \rho_s g+k_{gs}\left(v_g-v_s\right) + \sum_{n=1, s\neq n}^{N} k_{ns}\left(v_n-v_s\right)+\nabla\tau_s \tag{4}$$

Here, P_g, k_{gs}, k_{ns}, τ_g and g are the fluid pressure, gas-solid momentum exchange coefficient between the gaseous and solids phases, solid-solid momentum exchange coefficient, the viscous stress tensor of the gas phase and the gravity constant, respectively.

The Newtonian form of the viscous stress tensor for the gas phase, τ_g in Eq (3), and for the solids phase, τ_s in Eq (4), are given by (Liu et al., 2016):

$$\tau_g=v_g\left[\nabla v_g+\left(\nabla v_g\right)^T-\frac{2}{3}\mu_g\left(\nabla\cdot v_g\right)\right]I \tag{5}$$

$$\tau_s=\left(-P_s+\zeta_s\nabla\cdot v_s\right)I+\mu_s\left\{\left[\nabla v_s+\left(\nabla v_s\right)^T\right]-\frac{2}{3}\left(\nabla\cdot v_s\right)I\right\} \tag{6}$$

Here P_s, μ_s, ζ_s and I are the solids pressure, the solids viscosity, the solids bulk viscosity and the unit tensor, respectively.

P_s represents the solid pressure (normal forces) created due to particle-particle collisions in a flow due to presence of several solid phases (Gidaspow, 1994)

$$P_s=\varepsilon_s \rho_s \Theta_s+\sum_{n=1}^{N} 2\frac{d_{ns}^3}{d_s^3}\left(1+e_{ns}\right)g_{0,ns}\varepsilon_n \varepsilon_s \rho_s \Theta_s \tag{7}$$

e_{sn} is the particle-particle restitution coefficient between phase s and n. d_s is the particle diameter. d_{ns} is the mean diameter of the particles in phase n and s. $g_{o,ns}$ and Θ_s are the radial distribution function and the granular temperature respectively.

The bulk viscosity of the solids, ζ_s in Eq (6), is given by (Neri and Gidaspow, 2000):

$$\zeta_s=\frac{4}{3}\varepsilon_s^2 \rho_s d_s g_{0,ss}\left(1+e_{ss}\right)\sqrt{\frac{\Theta_s}{\pi}} \tag{8}$$

The solids shear viscosity in Eq (6) is given as (Arastoopour, 2001):

$$\mu_s=\frac{4}{5}\varepsilon_s^2 \rho_s d_s g_{0,ss}\left(1+e_{ss}\right)\sqrt{\frac{\Theta_s}{\pi}}+\frac{10\rho_s d_s\sqrt{\pi\Theta_s}}{96\left(1+e_{ss}\right)\varepsilon_s g_{0,ss}}\left[1+\frac{4}{5}\varepsilon_s g_{0,ss}\left(1+e_{ss}\right)\right]^2 \tag{9}$$

Wen and Ergun (Huilin and Gidaspow, 2003) proposed that the exchange coefficient k_{gs} between the gas and the solids phase given in Eq (4) and (5) could be calculated by:

$$k_{gs}\big|_{Wen\ \&\ Yu}=\frac{3}{4}C_D\frac{\rho_s\varepsilon_s\left|v_g-v_s\right|}{d_s}\varepsilon_g^{-2.65} \qquad \varepsilon_g>0.8 \tag{10}$$

$$k_{gs}\big|_{Ergun}=150\frac{\left(1-\varepsilon_g\right)\varepsilon_s\mu_g}{\left(\varepsilon_g d_s\right)^2}+1.75\frac{\rho_g\varepsilon_s\left|v_g-v_s\right|}{\varepsilon_s d_s} \quad \varepsilon_g\leq 0.8 \tag{11}$$

The drag coefficient depends on the value of the Reynolds number, Re:

$$\begin{cases} C_D=\dfrac{24}{\mathrm{Re}}\left(1+0.15\,\mathrm{Re}^{0.687}\right) & \mathrm{Re}<1000 \\ C_D=0.44 & \mathrm{Re}\geq 1000 \end{cases} \tag{12}$$

$$\mathrm{Re}=\frac{\rho_g\varepsilon_g\left|v_g-v_s\right|d_s}{\mu_g} \tag{13}$$

2.1.3 Kinetic Theory of Granular Flow

The continuum approach of granular flow requires a method to calculate viscous forces applied on the solid particles. Particle collisions are considered to predict the physical properties of the particulate phase. The theory is widely used in modelling of particulate systems to simulate actual systems with a high level of accuracy.

The kinetic theory introduces a new variable Θ, called granular temperature, and it is a measure of the kinetic energy of the solids. One-third of the mean square velocity of the random motion of the particles is considered as the granular temperature, $\Theta = v_s'^2/3$, where $v_s'^2$ is the square of the fluctuating velocity of the particle. A transport equation for the granular temperature can be written as (Huilin et al., 2001):

$$\frac{3}{2}\left[\frac{\partial}{\partial t}(\varepsilon_s\rho_s\Theta_s)+\nabla\cdot(\varepsilon_s\rho_s\Theta_s)v_s\right]=(\nabla PI+\varepsilon_s\nabla\tau_s):\nabla v_s+\nabla\cdot(k_s\nabla\Theta_s)-\gamma_s+\Phi_s+D_{gs} \quad (14)$$

Here, γ_s is dissipation of turbulent kinetic energy, Φ_s is energy exchange between gas and particle and D_{gs} is energy dissipation.

The turbulent kinetic energy dissipation, γ_s in Eq (14), is given as (Neri and Gidaspow, 2000):

$$\gamma_s = 3(1-e_{ss}^2)\varepsilon_s^2\rho_s g_{0,ss}\Theta_s\left(\frac{4}{d_s}\sqrt{\frac{\Theta_s}{\pi}}-\nabla\cdot v_s\right) \quad (15)$$

The radial distribution for N solid phases can be expressed as (Ahmadi and Ma, 1990):

$$g_{o,ss} = \frac{1+2.5\varepsilon_s+4.59\varepsilon_s^2+4.52\varepsilon_s^3}{\left(1-\left(\frac{\varepsilon_s}{\varepsilon_{s,\max}}\right)^3\right)^{0.678}}+\frac{1}{2}d_s\sum_{n=1}^{N}\frac{\varepsilon_n}{d_n} \quad (16)$$

$$\varepsilon_s = \sum_{n=1}^{N}\varepsilon_n \quad (17)$$

n are solid phases only, and d_s is the diameter of a particle in the s^{th} phase.

The energy exchange between the gas and the solids phases in Eq (14) is defined as:

$$\Phi_s = -3k_{gs}\Theta_s \quad (18)$$

The rate of energy dissipation per unit volume is expressed in the following equation:

$$D_{gs} = \frac{d_s\rho_s}{4\sqrt{\pi\Theta_s}}\left(\frac{18\mu_g}{d_s^2\rho_s}\right)^2|v_g-v_s|^2 \quad (19)$$

2.2 Simulation

The simulations were performed under a Froude number of 16×10^{-4} to maintain the rolling mode in the rotating drum. The cylinder and the particles rotate in the counterclockwise direction. The drum was simulated with different values for the degree of filling: 10, 15 and 20 % of the drum height.

2.2.1 Physical Properties of Materials and Model Parameters

Two granular phases of $CaCO_3$ and Al_2O_3 were used in the simulations. The physical properties of the solids, as well as the model parameters, are given in table 1.

Table 1. Physical Properties of Materials and Model Parameters

Parameter	*Description*	*Value*
ρ_{CaCO_3} (kg/m^3)	Particle density	1760
$\rho_{Al_2O_3}$ (kg/m^3)		3000
d_{CaCO_3} (μm)	Particle diameter	175
$d_{Al_2O_3}$ (μm)		1000
ω (rpm)	Rotational speed	2

2.2.2 Geometry and Mesh

A circular geometry with a diameter of 0.4 m was created with a mesh of 5500 elements to represent the transverse plane of the rotating cylinder. Figure 1 shows the mesh of the transverse plane.

2.2.3 Initial and Boundary Conditions

In rotating drums, particles are subjected to wall friction and gravity forces. A proper boundary condition should be used for the relative motion between the particles and the drum wall. Here, a no-slip condition was assumed, meaning that the relative velocities of the gas and the particles at the wall are set to zero.

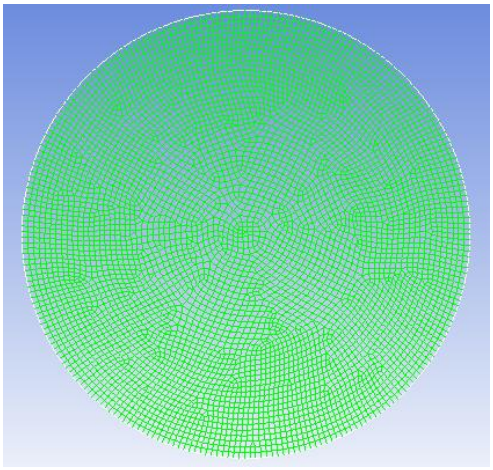


Figure 1. Mesh of the transverse plane

2.2.4 Solution Strategy and Convergence Criteria

In this study, the governing equations of the model were solved using the finite volume approach. The fluids were taken as incompressible, and a pressure-based solver was used. The "SIMPLE" algorithm (Patankar and Spalding, 1972) was used for the coupling between pressure and velocity. Discretization of the governing equations was carried out according to the second order upwind scheme (Banks and Henshaw, 2012). The "QUICK" scheme (Versteeg and Malalasekera, 2007) was used to discretize the volume fractions. Finally, the time step of the simulations was 10^{-3} s, and the residual values for convergence were set to 10^{-3}.

3 Results and Discussion

In rolling mode, particles in the active layer move relatively fast in comparison with the particles in the passive layer. The distance of the particle travel in the active layer is proportional to the square of the particle diameter (Boateng and Barr, 1996). Consequently smaller particles are likely to be concentrated in the mid-chord section.

3.1 Effect of Particle Size on Segregation

The effect of particle size on segregation of granular particles was examined by considering two particles with different size but with the same density. The simulation was carried out with particle sizes of 175 µm and 1000 µm. The density of the particles was the same as the density of $CaCO_3$. Figures 2 and 3 show a significant segregation due to particle size (percolation). The smaller particles concentrate in the mid-section of the moving bed while the larger particles sink towards the bottom of the drum.

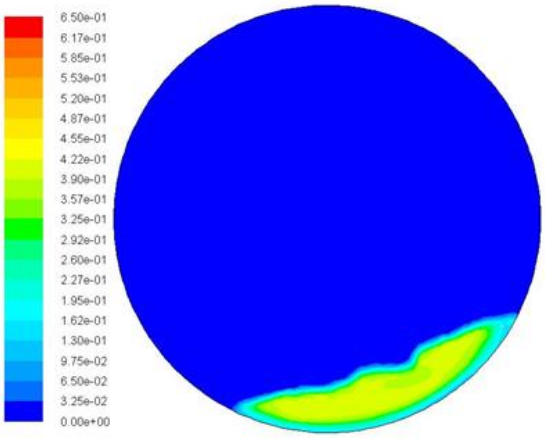


Figure 2. Volume fraction of particles with a diameter of 175 µm at pseudo steady-state.

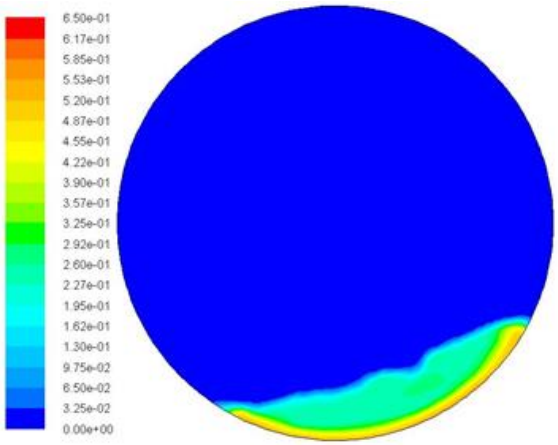


Figure 3. Volume fraction of particles with a diameter of 1000 µm at pseudo steady-state.

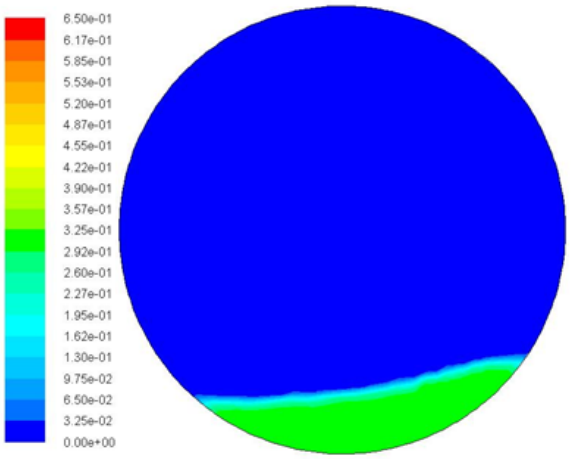


Figure 4. Volume fraction of particles with a density of 3000kg/m^3 in a mixture of particles with the same particle size at pseudo steady-state

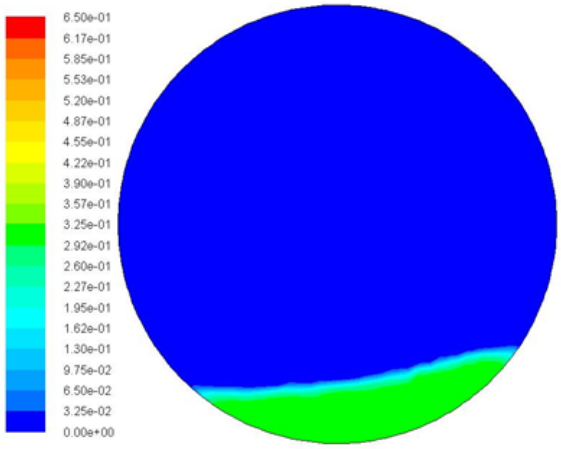
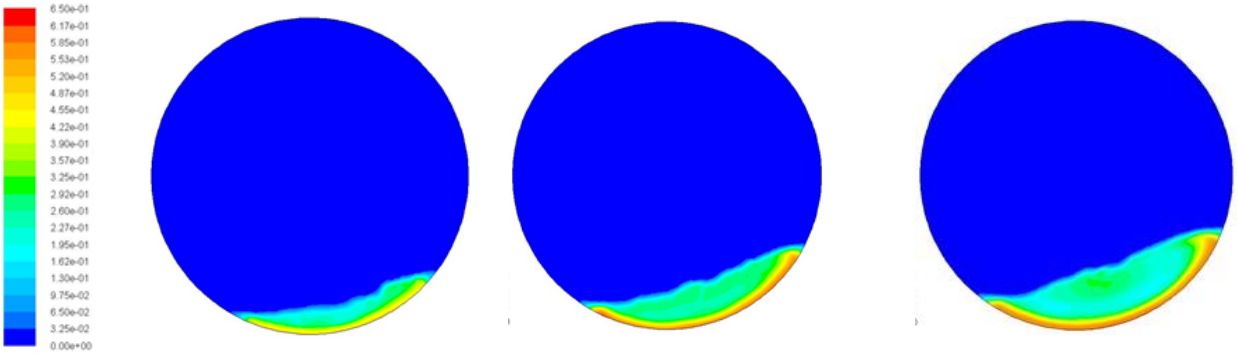
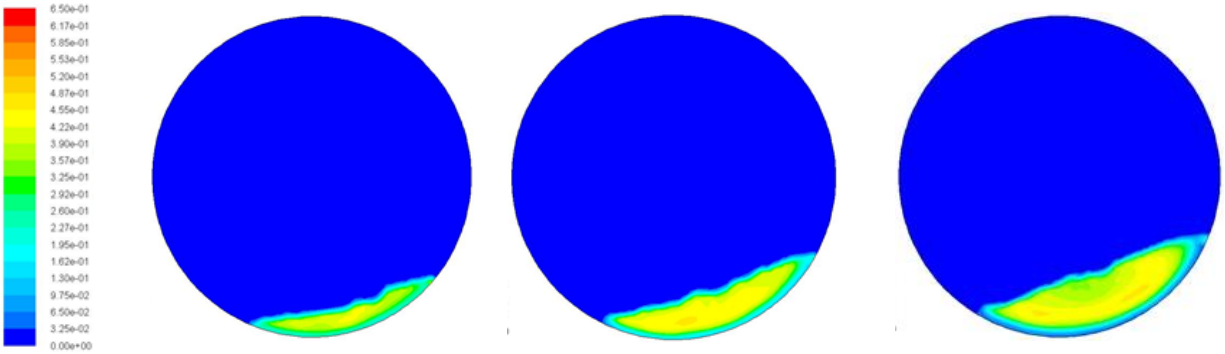


Figure 5. Volume fraction of particles with a density of 1760kg/m^3 in a mixture of particles with the same particle size at pseudo steady-state



a b c

Figure 6. Volume fraction of Al_2O_3 at pseudo steady-state with different degrees of particle filling, (a) 10%, (b) 15%, (c) 20%



a b c

Figure 7. Volume fraction of $CaCO_3$ at pseudo steady-state with different degrees of particle filling, (a) 10%, (b) 15%, (c) 20%

3.2 Effect of Density on Segregation

Segregation due to density variations was investigated considering the same particle size with different densities. The particle size was set to 1000 μm and the densities were 1760 kg/m^3 and 3000 kg/m^3. Volume fractions of the two particle types are shown in Figures 4 and 5, which show that for the particles considered no significant segregation appears due to density differences (condensation). According to Jha (2008), the particle size is the more dominant parameter for segregation, and this is in line with the present simulation results.

3.3 Effect of Degree of Filling on Segregation

Figures 6 and 7 show the volume fractions of $CaCO_3$ and Al_2O_3, respectively, with different degrees of fillings. The simulation results show that, with the selected particle sizes and densities, there will be a considerable segregation in all three cases. The segregation tendency is, however, more or less the same for all three, hence the degree of filling does not make a big impact on the particle motion pattern as such.

4 Conclusions

Two-dimensional simulations of the transverse plane in a rotating drum, based on the Eulerian approach and the kinetic theory of granular flow, indicate that particle segregation occurs mainly due to differences in particle size when the flow is in the rolling mode. The active layer in the rolling mode constitutes an environment facilitating segregation. Trajectory segregation and percolation seem to be the predominant mechanisms in the current simulations as no significant segregation appeared with different densities, hence segregation due to condensation was less important. The degree of particle filling in the rotary drum made no significant impact on the particle behavior.

References

G. Ahmadi and D. Ma. A thermodynamical formulation for dispersed multiphase turbulent flows—1. *International Journal of Multiphase Flow*, 16: 323-340, 1990.

H. Arastoopour. Numerical simulation and experimental analysis of gas/solid flow systems: 1999 Fluor-Daniel Plenary lecture. *Powder Technology*, 119: 59-67, 2001.

M. Azadi. Multi-fluid Eulerian modeling of limestone particles elutriation from a binary mixture in a gas solid fluidized bed. *Journal of Industrial and Engineering Chemistry*, 17: 229-236, 2011.

J. W. Banks and W. D. Henshaw. Upwind schemes for the wave equation in second-order form. *Journal of Computational Physics*, 231: 5854-5889, 2012.

A. A. Boateng, *Rotary Kilns: Transport Phenomena and Transport Processes.* USA: Butterworth-Heinemann publications, 2008.

A. A. Boateng and P. V. Barr. Modelling of particle mixing and segregation in the transvers plane of a rotary kiln. *Chemical Engineering Science,* 51: 4167-4181, 1996.

H. Chen, X. Zhao, Y. Xiao, Y. Liu, and Y. Liu. Radial mixing and segregation of granular bed bi-dispersed both in particle size and density within horizontal rotating drum. *Transactions of Nonferrous Metals Society of China,* 26: 527-535, 2016.

Y. Demagh, Hocine, B. Moussa, M. Lachi, and L. Bordja. Surface particle motion in rotating cylinders: Validation and similarity for an industrial scale kiln. *Powder Technology*, 224: 260-272, 2012.

Y. L. Ding, J. P. K. Seville, R. Forster, and D. J. Parker. Solid motion in rolling mode rotating drums operated at low to medium rotational speeds. *Chemical Engineering Science*, 56: 1769-1780, 2001.

G.J.Finnie, N.P.Kruyt, M.Ye, C.Zeilstra, and J.A.M.Kuipers. Longitudinal and transverse mixing in rotary kilns: A discrete element method approach. *Chemical Engineering Science*, 60: 4083-4091, 2005.

D. Gidaspow, *Multiphase flow and fluidization*. California: Academic press, 1994.

H. Henein, J. K. Brimacombe, and A. P. Watkinson. Experimental study of transverse bed motion in rotary kiln. *Metallurgical Transactions B,* 14B: 191-205, 1983.

L. Huilin and D. Gidaspow. Hydrodynamics of binary fluidization in a riser: CFD simulation using two granular temperatures. *Chemical Engineering Science,* 58: 3777-3792, 2003.

L. Huilin, D. Gidaspow, and E. Manger. Kinetic theory of fluidized binary granular mixtures. *Phys. Rev.* E, 64:061301: 1-8, 2001.

A. K. Jha. *Percolation segregation in multi-size and multi-component particulate mixtures: Measurement, sampling, and modeling*. PhD, Graduate School, College of Engineering Pennsylvania State University, 2008.

H. Liu, H. Yin, M. Zhang, M. Xie, and X. Xi. Numerical simulation of particle motion and heat transfer in a rotary kiln. *Powder Technology*, 287: 239-247, 2016.

A. Neri and D. Gidaspow. Riser hydrodynamics: Simulation using kinetic theory. *AIChE Journal*, 46: 52-67, 2000.

S. V. Patankar and D. B. Spalding. A calculation procedure for heat, mass and momentum transfer in three-dimensional parabolic flows. *International Journal of Heat and Mass Transfer*, 15: 1787-1806, 1972.

M. A. I. Schutyser, J. T. Padding, F. J. Weber, and W. J. Briels. Discrete particle simulations predicting mixing behavior of solid substrate particles in a rotating drum fermenter. *Inc. Biotechnol Bioeng*, 75: 666-675, 2001.

H. K. Versteeg and W. Malalasekera, *An introduction to computational fluid dynamics*, second ed. England: Pearson Education Limited 2007.

H. Yin, M. Zhang, and H. Liu. Numerical simulation of threedimensional unsteady granular flows in rotary kiln. *Powder Technology*, 253: 138-145, 2014.

Perspectives on Industrial Optimization based on Big Data Technology and Soft Computing through Image Coding

Yukinori Suzuki

Information and Electronic Engineering, Muroran Institute of Technology, Japan,
yuki@epsilon2.csse.muroran-it.ac.jp

Abstract

Industrial systems are being rapidly innovated due to recent information technology of IoT and fruitful results of artificial intelligence. We discuss roles of big data technologies and soft computing to optimize industrial systems and to design robust systems through image coding. We show a code book (CB) design for vector quantization (VQ) to discuss roles of soft computing and big data technology. The CBs were designed by conventional clustering algorithms. However, these conventional algorithms cannot provide CBs that encode and/or decode images with high image quality and low bits rate. We show a perspectives to overcome this problem to integrate big data technology and soft computing.

Keywords: industrial optimization, big data, soft computing, image coding

1 Introduction

Industrial systems and products are being rapidly innovated due to recent information technology and fruitful results of artificial intelligence. Nowadays, industrial systems are managed using the Internet, which is "the Internet of Things (IoT)". The development of products has therefore been dramatically speeded up, and industrial systems have also become both complex and large-scaled. To develop such systems, it is necessary to make them efficient for saving energy, downsizing, and reducing costs. Environmental compatibility is also important for developing the industrial systems without slowing down their performance as shown in Fig. 1. To design industrial systems satisfying the above requirements, optimization based on meta-huristic methods is a key point.

Many decision variables and objectives are involved in the design of industrial systems. Zhou et al. stated the problems on the many decision variables and objectives as follows (Zhou, 2014). (i) The correlation between the decision variables may be nonlinear, and some objectives are in conflict with each other. (ii) In meta-heuristic methods, since one population has a trade-off relationship and another population has a different trade-off relationship, there is conflict among populations. (iii) A multi-objective meta-heuristic method does not work efficiently when the number of objects is much larger than three, because Pareto-optimal solutions become intractable. (iv) Furthermore, when the number of objects increase, computational cost to obtain optimal solutions increases drastically. They also stated that it is necessary to develop optimization algorithms that can gain problem-specific knowledge during the optimization process to overcome the problems. If there is a large number of decision variables and objects, such knowledge is essential to focus the search in a promising direction. Big data technologies can provide us with such problem-specific knowledge. If we can obtain problem-specific knowledge using big data technology and provide a search direction for meta-huristic optimization, we may be able to obtain optimal solutions efficiently (Zhou, 2014).

For large-scaled complex systems, there are large amounts of uncertainties and impreciseness. They are involved by varying environmental conditions, system degenerations, or changing customer demand (Zhou, 2014). Furthermore, to optimize complex systems, the principle of incompatibility suggested by Zadeh is essential. The principle states that as the complexity of a system increases beyond a threshold, precision and significance become almost exclusive (Zadeh, 1973). Methodologies comprising soft computing (SC) provide an approximate and adequate solution for uncertainties, impreciseness, and problems caused by complexity such as nonlinearity and non-stationrity (Suzuki, 2013). SC principally consists of fuzzy logic, evolutionary computation, neural networks, probabilistic computing, and a rough set. These methodologies are not exclusive but are complementary. To obtain the optimal solution efficiently and treat uncertainties related to complex industrial systems, it is necessary to integrate big data technology and soft computing as shown in Fig. 2. In this paper, we discuss roles of big data technology and soft computing through image coding. Furthermore, we give perspectives about integration of big data technology and soft computing.

We have been studying vector quantization (VQ) for image coding (Sasazaski, 2008; Miyamoto, 2010). Fig. 3 shows a conceptual diagram of VQ. For VQ, an image is divided into blocks of pixels such as 4×4 or 8×8. Each block of the image is encoded using a CB, which consists of code vectors (CVs). The nearest CV in the CB is taken and its index is memorized in the index map. The indexes are transferred to the destination through a communication channel to decode images. In decoding,

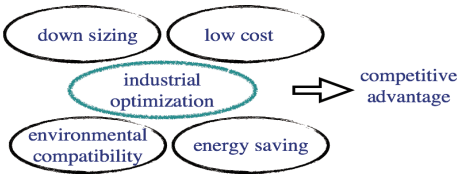


Figure 1. Industrial systems and optimization technology.

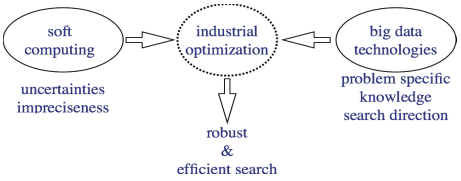


Figure 2. Integration of big data technology and soft computing for optimization.

the CV corresponding to an index is retrieved from the CB to reconstruct the image. Since a CB determines the performance of image coding with VQ, design of a CB is essential for VQ. Various type of images have to be encoded for sending and have to be decoded for receiving. Since a huge number of images is being transmitted through communication channels, we cannot predict the images to be encoded and/or decoded before VQ. We have to design a CB that can encode and/or decode images not only with maintenance of high image quality but also at a high compression rate. There is an enormous amount of image data in cyberspace. A huge number of people release their photographs in websites and also there are huge image databases. A flood of images satisfies the definition of big data of three “Vs": volume (large datasets), variety (different types of data from myriad sources), and velocity (data collected in real time) (Fang, 2015). We use big data of images to design a CB that is able to encode and/or decode a variety of images as shown Fig. 4. We acquire big data of images from the database and analyze them to extract their features. These features are grouped

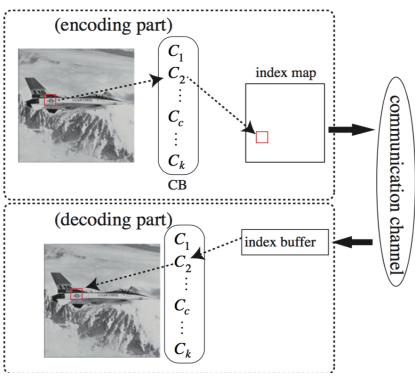


Figure 3. Conceptual diagram of vector quantization.

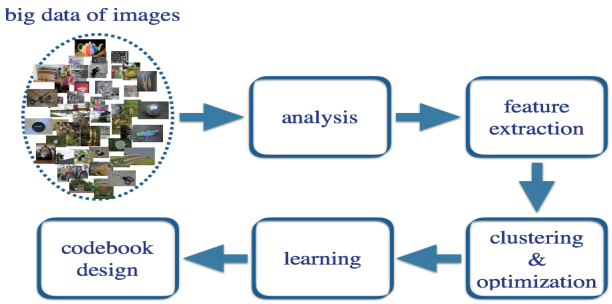


Figure 4. Big data technology and soft computing to design a CB.

into categories using a clustering algorithm and optimization techniques. The categorized features are obtained by a CB using learning algorithms. The CB designed in this framework can be expected to encode and/or decode images with good quality and high compression rate.

In the following section, we show the conventional CB design methods using clustering algorithms. Four widely used clustering algorithms were used. As crisp clustering algorithm k means clustering (KMC) and the enhance LBG (ELBG) clustering algorithm were used. As fuzzy clustering algorithms, fuzzy k means (FKM) clustering and fuzzy learning vector quantization (FLVQ) algorithms were used. In section 3, computational experiments to evaluate four the clustering algorithms are shown. The roles of big data technology and soft computing are also discussed and surveyed. Finally, the present paper is concluded in section 4.

2 Clustering Algorithms

2.1 k Means Clustering Algorithm

The k means clustering (KMC) algorithm is the most commonly used algorithm due to its algorithmic simplicity and low computational cost. It is also called the LBG algorithm (Linde, 1980). There are two methods of initialization for the LBG algorithm: random initialization and initialization by splitting. CB initialization is a very important task. For the KMC algorithm, we chose random initialization. The KMC algorithm assigns each input vector to a certain cluster. As initial centroids (cluster centers), the number of c input vectors is chosen. NNC (nearest neighbor condition) and CC (central condition) are used for optimal clustering. For NNC, input vector $\mathbf{x_i}$ is assigned to the cth cluster when $d(\mathbf{x}_i, \mathbf{c}_c) = \min_{\mathbf{c}_c \in C} d(\mathbf{x}_i, \mathbf{c}_c)$ is satisfied. We employed the squared Euclidean norm for clustering such as $d(\mathbf{x}_i, \mathbf{c}_c) = \|\mathbf{x}_i - \mathbf{c}_c\|^2$. After all input vectors have been assigned to respective centroids, new centroids are updated according to CC. In the KMC algorithm, since the input vector is assigned to only one centroid, which is crisp clustering, the membership function

takes zero or one.

$$u_c(\mathbf{x}_i) = \begin{cases} 1 \text{ if } d(\mathbf{x}_i, \mathbf{c}_c) = \min_{\mathbf{c}_c \in C} d(\mathbf{x}_i, \mathbf{c}_c) \\ 0 \qquad\qquad\qquad\qquad \text{otherwise.} \end{cases} \tag{1}$$

Once we obtain the membership function, the centroids are updated according to

$$c_c = \frac{\sum_{i=1}^{n} u_c(\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n} u_c(\mathbf{x}_i)}, \tag{2}$$

where $c = 1, \cdots, k$. Iterative updating the membership functions and centroids by (1) and (2) minimize MQE estimated as

$$MQE = \frac{1}{n} \sum_{c=1}^{k} \sum_{i=1}^{n} u_c(\mathbf{x}_i) \|\mathbf{x}_i - \mathbf{c}_c\|^2. \tag{3}$$

We conclude the KMC algorithm as follows: (i) The number of k input vectors is randomly selected as initial centroids, (ii) the membership function is computed using (1) for all input vectors, (iii) after new membership functions have been obtained, all centroids are updated according to (2). (ii) and (iii) are repeated as long as the convergence condition to terminate repetition is not satisfied. In this paper, the convergence condition is determined as

$$\frac{|MQE(\upsilon - 1) - MQE(\upsilon)|}{MQE(\upsilon)} < \varepsilon, \tag{4}$$

where υ is the number of iterations and $\varepsilon = 10^{-4}$.

2.2 Enhanced LBG Algorithm

Patane et al. (Patane, 2001) pointed out that there are two important drawbacks of the LBG algorithm. One drawback is an empty cluster that is generated when all input vectors are nearer to other CVs. This empty cluster is generated due to inappropriate selection of the initial CVs. As for the other drawback, suppose that there are two clusters and three CVs. In the smaller cluster, there are two CVs. However, there is one CV in the larger cluster. All input vectors in the smaller cluster are approximated well by the two CVs, but the input vectors in the larger cluster are poorly approximated by the CV. For optimal clustering, the larger cluster should include two CVs, while the smaller cluster includes one CV. However, it is impossible to implement CV migration for the LBG algorithm. Patane et al. (Patane, 2001) claimed that this impossible migration is a great limitation of the LBG algorithm. To improve the performance of the LBG algorithm, they developed a migration algorithm for the LBG algorithm that called the enhanced LBG algorithm.

Patane at al. (Patane, 2001) introduced a quantity of the utility of CVs, which provides a solution to overcome the drawbacks stated above. The utility index of the cth cluster is computed as

$$D_{mean} = \frac{1}{n} \sum_{c=1}^{k} D_c, \tag{5}$$

where D_c is the distortion value of the cth cluster. The utility index of the kth cluster is

$$U_c = \frac{D_c}{D_{mean}}, c = 1, \cdots, k. \tag{6}$$

Migration of CVs from a smaller cluster to a larger cluster is implemented using the utility of CVs. The algorithm was named ELBG block, in which the utility of each cluster is estimated and clusters with low utility are found. The algorithm could implement the “partial distortion" theorem by Gersho (Gersho, 1979).

2.3 Fuzzy k Means Clustering Algorithm

The idea of fuzzy sets was introduced to allow multiple assignments of input vectors to CVs, which is implemented by the fuzzy k means clustering algorithm (FKM). This algorithm is an extension of the KMC algorithm using fuzzy sets (Bezdek, 1987). To derive the FKM algorithm, NNC and CC conditions were used to design an optimal clustering algorithm. The membership function is the degree of belongness of the input vector to a certain cluster. It is determined so as to minimize total distortion

$$D_{total} = \sum_{c=1}^{k} \sum_{i=1}^{n} u_c(\mathbf{x}_i)^m \|\mathbf{x}_i - \mathbf{c}_c\|^2 \tag{7}$$

under the two constraints

$$0 < \sum_{i=1}^{n} u_c(\mathbf{x}_i) < n \tag{8}$$

$$\sum_{c=1}^{k} u_c(\mathbf{x}_i) = 1, \tag{9}$$

where $1 < m < \infty$ provides the fuzziness of the clustering. In the case of $m = 1$, FKM clustering becomes crisp clustering and m has to be given in advance. The membership function is computed as

$$u_c(\mathbf{x}_i) = \frac{1}{\sum_{l=1}^{k} \left(\frac{d(\mathbf{x}_i, \mathbf{c}_c)}{d(\mathbf{x}_i, \mathbf{c}_l)} \right)^{\frac{2}{m-1}}}, \tag{10}$$

where $d(\mathbf{x}_i, \mathbf{c}_c)$ and $d(\mathbf{x}_i, \mathbf{c}_l)$ are the squared Euclidean norm. If the norm became zero, it was replaced by one to avoid zero division in our experiments. After the membership function has been updated, the centroids are renewed according to the CC condition.

$$\mathbf{c}_c = \frac{\sum_{i=1}^{n} u_c(\mathbf{x}_i)^m \mathbf{x}_i}{\sum_{i=1}^{n} u_c(\mathbf{x}_i)^m}. \tag{11}$$

Karayiannis et al. (Karayiannis, 1995) reported that the FKM algorithm showed the best performance when $m = 1.2$, and we therefore used this value for our experiments. The convergence condition to terminate the repetition was the same as that in (4). If the repetition was more than 500, we made the repetition terminate in the experiment.

2.4 Fuzzy Learning Vector Quantization

Tsao et al. (Tsao, 1994) developed a clustering algorithm with integration of FKM and KMC based on Kohonen's learning vector quantization (LVQ) algorithm. This algorithm is fuzzy learning vector quantization (FLVQ). In FLVQ, transition from fuzzy mode to crisp mode is implemented by controlling the fuzziness parameter (Tsekouras, 2008). This means transition from assignment of multiple clusters to assignment of a single cluster. The objective function to minimize distortion and constraints during repetition of the FLVQ algorithm is the same as that in (7), (8), and (9). According to (Tsekouras, 2008), the FLVQ algorithm consists of the following stages.

(stage 1) Specify the number of clusters k and the initial CVs

$c_1, c_2, \cdots c_k$.

(stage 2) Set the maximum number of iterations t_{max}, the initial m_0 and the final m_f values for the fuzziness parameters.

(stage 3) for $t = 0, 1, 2, \cdots, t_{\max}$

(i) The fuzziness parameter $m(t) = m_0 - \left[t(m_0 - m_f)\right] / t_{\max}$ is computed.

(ii) The membership function is updated as

$$u_c(\mathbf{x}_i) = \frac{1}{\sum_{l=1}^{k} \left(\frac{d(\mathbf{x}_i, \mathbf{c}_c)}{d(\mathbf{x}_i, \mathbf{c}_l)} \right)^{\frac{2}{m-1}}}, \tag{12}$$

where m is $m(t)$. We used the squared Euclidean norm for $d(\mathbf{x}_i, \mathbf{c}_c)$ and $d(\mathbf{x}_i, \mathbf{c}_l)$. If the norm became zero, it was replaced by one to avoid zero division in our experiments.

(iii) The CVs are updated using new membership functions as

$$\mathbf{c}_c = \frac{\sum_{i=1}^{n} u_c(\mathbf{x}_i)^m \mathbf{x}_i}{\sum_{i=1}^{n} u_c(\mathbf{x}_i)^m}, \tag{13}$$

where m is $m(t)$.

(iv) The repetition terminates when the following condition is satisfied:

$$\sum_{c=1}^{k} \left\| \mathbf{c}_\mathbf{c}^{\mathbf{t-1}} - \mathbf{c}_\mathbf{c}^{\mathbf{t}} \right\|^2 < \varepsilon, \tag{14}$$

where $\varepsilon = 10^{-4}$.

We used the following parameters for our experiments: $m_0 = 1.5$, $m_f = 1.001$ and $t_{max} = 100$.

3 Performance of Clustering Algorithms and A Role of Big Data Technology

We choose images and design a CB using those images to encode and/or decoded images by VQ in advance. In the set of experiments, we examined the importance of selection of clustering algorithms for designing a CB. Performance of clustering algorithms was estimated. We carried out sets of computational experiments using 8-bit gray scale images (images used in the experiments were images from CVG-UGR-Image database http://decsai.ugr.es/cvg/dbimagenes/index.php). In the set of experiments, two sizes of images were used: 256 × 256 pixels and 512 × 512 pixels. The images used as both the learning images and test images are shown in Fig. 5. These images were segmented into 4×4 pixels as a block of pixels. Each block of pixels was treated as a learning and test vector with 16 dimensions. When the image is 256 × 256 pixels, there are 4096 learning and test vectors. These 4096 learning vectors were used to de-

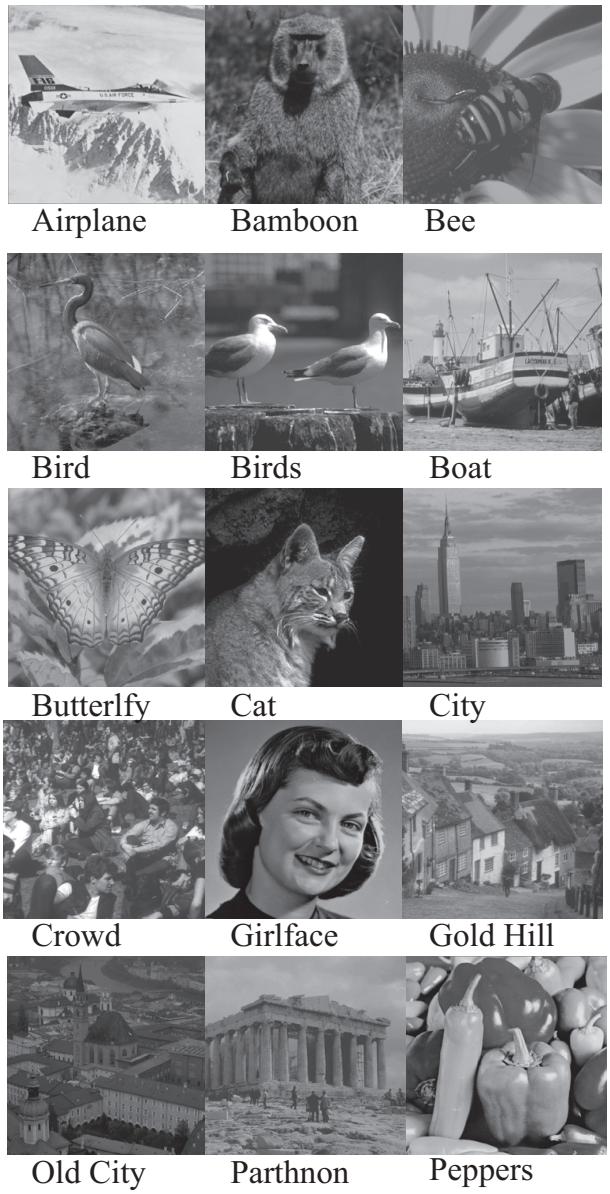


Figure 5. Images used as both learning and test images. There are two sizes for each image: 256×256 pixels and 512×512 pixels.

sign a CB using four clustering algorithms (KMC, ELBG, FKM, FLVQ). For example, the image “Airplane" was segmented into 4096 blocks as both learning and test vectors. Four CBs were designed using these learning vectors with four clustering algorithms. We decoded test vectors using the four CBs, and the performance of the clustering algorithms was estimated by image quality of the decoded image. Image quality is estimated in terms of *PSNR*.

$$PSNR = 10\log_{10}\left(\frac{PS^2}{MSE}\right)(\mathrm{dB}), \tag{15}$$

where $PS = 255$. *MSE* is the mean square error between the original image and the decoded image. We performed experiments with increases in the number of CVs as 64, 128, 256, 512, and 1024.

Fig. 6 shows a comparison of the performance of the four clustering algorithms. Each value is the average of *PSNR*s for 15 test images. As shown in Fig. 6, the ELBG algorithm showed the best performance for both image sizes of 256×256 and 512×512 pixels. The difference between the *PSNR* of the ELBG algorithm and the *PSNR*s of the other algorithms increases as the number of CVs increases. The performance of the FKM algorithm and that of the FLVQ algorithm were comparable for both image sizes. The KMC algorithm showed poor performance in comparison with the performance of the ELBG, FKM, FLVQ algorithms. When the number of CVs was 256 and image size was 256×256, the average *PSNR*s of the clustering algorithms were 30.26 dB (KMC), 31.29 dB (ELBG), 30.67 dB (FKM), and 30.79 dB (FLVQ). The difference between the *PSNR*s of the KMC and ELBG algorithms was 1.03 dB, which is sufficient to perceive a difference. Fig. 7 and Fig. 8 show decoded images with CBs (the number of CVs being 256) constructed by the KMC and ELBG algorithms, respectively. In Fig. 7, the upper image is the original “Airplane" image. The middle image was decoded using CBs designed by the KMC algorithm, and the bottom image was decoded using CBs designed by the ELBG algorithm. The difference between *PSNR*s in images decoded by CBs constructed by the KMC and ELBG algorithms was 1.63 dB. The difference between *PSNR*s was 1.17 dB in the case of “Girlface" in Fig. 8. We can perceive a difference in decoded image quality between the middle and bottom images. For example, in the “Airplane" image in the middle panel, it is difficult to recognize the letters on the tail. However, we can clearly perceive the letters on the tail in the image in the bottom panel. In the image of “Girlface", there are strong block noises at the lower jaw and lip in the image in the middle panel. Block noises at the lower jaw and lip are decreased in the image in the bottom panel. In the case of image size being 512×512, the average *PSNR*s of the clustering algorithms were 31.22 dB (KMC), 31.67 sB (ELBG), 31.41 dB (FKM), and 31.42 dB (FLVQ). The difference in *PSNR* between KMC and ELBG was 0.45 dB, which enables us to perceive an image difference. The experiments demonstrated that selection of a clustering algorithm is important to design a CB.

Table 1. Compression rates when segmentation block size is 4×4. Overhead in bits/pixel to transmit CB was neglected. In the table, CVs shows the number of CVs.

CVs	64	128	256	512	1024
bits/pixel	0.375	0.4375	0.5	0.5628	0.625

When segmented block size is 4×4, the compression rate is given as shown in Table 1. In practical application of VQ, it is necessary to keep compression rate less than 0.5 bits/pixel. Furthermore, we consider that it is desiable to keep image quality (*PSNR*) more than 35.0 dB. If *PSNR* of the compressed image is more than 35.0 dB, we cannot perceive the difference between original image and compressed image. However, *PSNR*s of the compression images are slightly more than 30.0 dB in the case of image being 256×256 as shown in Fig. 6. ELBG algorithm showed the best performance (*PSNR* was 31.29 dB). In the case that image was 512×512, *PSNR*s of decoded image were slightly more than 31.0 dB. ELBG algorithm also showed the best performance of 31.67. These values are far from 35.0 dB. In this sense, conventional CB design by clustering could not be applicable to encoding and/or decoding images. To overcome this limitation, we use big data technology as shown in Fig. 4. We collect a huge number of images from cyberspace or a database and extract the features of the images collected. The features of the images are categorized using a clustering algorithm and they are optimized to select essential features. The selected features are learned by neural networks (soft computing) to construct a CB. This is the author’s perspective and opinion for designing a CB based on big data technology. So far, we have no evidence showing the performance of the CB designed by the method described above. We intend to evoke discussion rather than to provide evidences of big data technology for image coding.

4 Conclusions

Roles of big data technology and soft computing for industrial optimization were discussed in this paper. It is necessary to optimize industrial systems for saving energy, downsizing, and reducing cost. Image coding by VQ was presented to discuss the necessity of big data technology and soft computing. CBs to encode and/or decode images were designed using conventional clustering algorithms. However, performance of the CBs designed by conventional clustering algorithms did not show decoded image quality more than 32 dB of average *PSNR* when the number of CVs is 256. This image quality is not sufficient for practical application of image coding. We therefore showed a perspective using big data technology and soft computing for discussion.

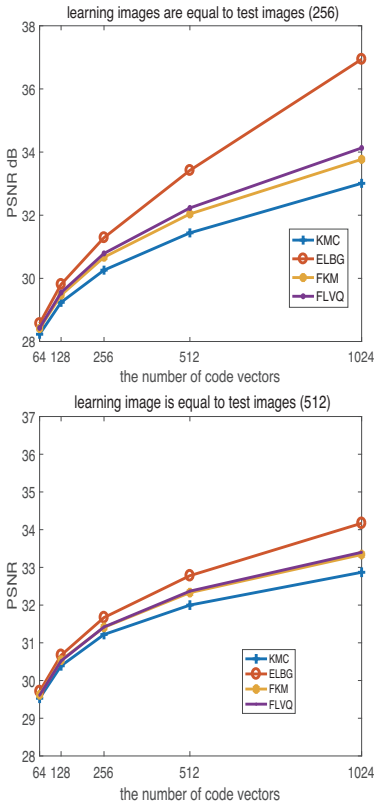


Figure 6. Changes in *PSNR*s with increases in the number of CVs . *PSNR* values in the graph were average values for 15 decoded test images. The upper panel shows *PSNR*s when image size was 256×256 pixels. The lower panel shows *PSNR*s when image size was 512×512 pixels.

Acknowledgment

This research project was partially supported by a grant-in-aid for scientific research from the Japan Society for Promotion of Science (15K00329).

References

Z. Zhou, N.V. Chawla, Y. Jin, and G.J. Willians. Big data opportunities and challenges: discussions from data analytic perspectives. *IEEE computational intelligence Magazine*, 9(4): 62-74, 2014.

L. A. Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans. Systems, Man, and Cybernetics*,3(1):24-44, 1973.

Y. Suzuki. Optimization, data mining and industrial applications using soft computing. *Proceedings of the Joint Seminar Finland and Japan*, 1-8, 2013.

K. Sasazaki, S. Saga, J. Maeda, and Y. Suzuki. Vector quantization of images with variable block size. *Applied Soft Computing*, 8(1):634-645, 2008.

H. Matsumoto, F. Kichikawa, K. Sasazaki, J. Maeda, and Y. Suzuki. Image compression using vector quantization with variable block size division. *IEEJ Trans EIS*, 130(8):1431-1439, 2010.

H. Fang, Z. Zhang, C.J. Wang, M. Daneshmand, C. Wang, and H. Wang. A survey of big data research. *IEEE Network*, 29(5):6-9, 2015.

Y. Linde, A. Buso, and R. Gray. An algorithm for vector quantization design. *IEEE Trans Communications*, 28(1):84-94, 1980.

G. Patane and M. Russo. The enhanced LBG algorithm. *Neural Networks*, 14(9):1219-1237, 2001.

A. Gersho. Asymptotically optimal block quantization. *IEEE Trans. Information Theory*, 25(4):373-380, 1979.

J.C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, 1987.

N.B. Karayiannis and P. Pai. Fuzzy vector quantization algorithms and their application in image compression. *IEEE Trans. Image Processing*, 4(9):1193-1201, 1995.

E. C-K Tsao, J. C. Bezdek, and N.R. Pal. Fuzzy Kohonen clustering networks. *Pattern Recognition*. 27(5):757-764, 1994.

G. E. Tsekouras, M. Antonios, C. Anagnostopoulos, D. Gavalas, and D. Economou. Improved batch fuzzy learning vector quantization for image compression. *Information Sciences*, 178(20):3895-3907, 2008.

D. Tsolakis, G.E. Tsekouras, and J. Tsimikas. Fuzzy vector quantization for image compression based on competitive agglomeration and a novel codeword migration strategy. *Engineering applications of artificial intelligence*, 25(6):1212-1225, 2012.

Figure 7. Upper image is the original “Airplane" image. Middle and bottom images were decoded using CBs designed by KMC and ELBG algorithms, respectively. These images are 256×256 pixels in size.

Figure 8. Upper image is the original “Girlface" image. Middle and bottom images were decoded using CBs designed by KMC and ELBG algorithms, respectively. These images are 256×256 in size.

A Novel Metaheuristic Algorithm inspired by Rhino Herd Behavior

Gai-Ge Wang[1] Xiao-Zhi Gao[2] Kai Zenger[2] Leandro dos S. Coelho[3]

[1]School of Computer Science and Technology, Jiangsu Normal University, China, gaigewang@163.com
[2]Department of Electrical Engineering and Automation, Aalto University, Finland,
xiao.z.gao@gmail.com, kai.zenger@aalto.fi
[3]Industrial and Systems Engineering Graduate Program, University of Parana, Brazil,
leandro.coelho@pucpr.br

Abstract

In this paper, inspired by the herding behavior of rhinos, a new kind of swarm-based metaheuristic search method, namely Rhino Herd (RH), is proposed for solving global continuous optimization problems. In various studies of rhinos in nature, the synoptic model is used to describe rhino's space use and estimate its probability of occurrence within a given domain. The number of rhinos increases year by year, and this increment can be forecasted by several population size updating models. Synoptic model and a population size updating model are formalized and generalized to a general-purpose metaheuristic optimization algorithm. In RH, null model without introducing any influences is generated as the initial herding. This is followed by rhino modification via synoptic model. After that, the population size is updated by a certain population size updating model, and newly-generated rhinos are randomly initialized within the given conditions. RH is benchmarked by fifteen test problems in comparison with biogeography-based optimization (BBO) and stud genetic algorithm (SGA). The results clearly show the superiority of RH in searching for the better function values on most benchmark problems over BBO and SGA.

Keywords: rhino herd, synoptic model, population size updating model, benchmark functions, swarm intelligence

1 Introduction

The current real-world optimization problems are increasingly more and more complex and they are hard to be solved by the traditional mathematical methods. On the other hand, human beings are always learning the rule of nature, and improve the ability to handle the complicated problems. By learning the collective behavior of systems, swarm intelligence (SI) (Cui and Gao, 2012) is studied.

Since they are put forward, SI-based algorithms are becoming more and more popular in several engineering applications because of their promising performances when addressing different kinds of real-world optimization problems, such as test-sheet composition (Duan et al., 2012), target threat assessment (Wang et al., 2012a), parameter estimation (Li and Yin, 2014), feature selection (Li and Yin, 2013a), path planning (Wang et al., 2016a; Wang et al., 2012b), wind generator design (Gao et al., 2012a,b), nonlinear system modeling (Gandomi and Alavi, 2011), scheduling (Li and Yin, 2013b), neural network training (Mirjalili et al., 2014a) and knapsack problem (Zou et al., 2011; Feng et al., 2017; Feng et al., 2014). Although SI algorithms involve a great number of methods, particle swarm optimization (PSO) (Kennedy and Eberhart, 1995; Mirjalili et al., *2014b;* Wang et al., *2016b;* Wang et al., *2014c;* Zhao et al., *2012,* Zhao, *2010;* Mirjalili et al., *2013)* and ant colony optimization (ACO) (Dorigo et al., *1996)* are two of the most representative and widely used ones so far. They are inspired by the social behavior of bird when searching for food and remembering paths via pheromone. Recently, inspired by swarm behavior of different animals, serials of SI algorithms have been developed and proposed, such as artificial bee colony (ABC) (Karaboga and Basturk, 2007), elephant herding optimization (EHO) (Wang et al., 2015a;Wang et al., *2016b)* chicken swarm optimization (CSO) (Meng et al., 2014) bird swarm algorithm (BSO) (Meng et al., 2015) cuckoo search (CS) (Yang and Deb, 2009; Li et al., 2013; Wang et al., 2016c; Wang et al., *2016d;*Wang et al., *2016e ;*Li and Yin, *2015)* bat algorithm (BA) (Yang, 2010; Mirjalili et al., 2013; Zhang and Wang, 2012; Wang et al., 2015b), firefly algorithm (FA) (Gandomi et al.,2011; Yang, 2010; Wang et al., *2014d;* Guo et al., *2013)* ant lion optimizer (ALO) (Mirjalili, 2015), chaotic swarming of particles (CSP) (Kaveh et al., 2014) monarch butterfly optimization (MBO) (Wang et al., 2015c;Wang et al., *2016e;*Wang et al., *2016f;*Ghetas et al., *2016)* krill herd (KH) (Gandomi and Alavi, 2012; Wang et al., 2013; Gandomi et al., *2013;* Wang et al., *2014d;*Wang et al., *2014e;* Wang et al., *2014f;*Wang et al., *2016f;*Guo et al., *2014;* Wang et al., *2016g;*Li et al., *2015)* multi-verse optimizer (MVO) (Mirjalili et al., 2016) dragonfly algorithm (DA) (Mirjalili, 2016), and grey wolf optimizer (GWO) (Mirjalili et al.,.2014; Saremi et al., *2014)* These algorithms have been successfully used to address an array of real-world problems.

Except for SI algorithms, inspired by the evolutionary rule of nature, evolutionary algorithms (EAs) are proposed. Among different kinds of EAs, the following algorithms are some of the most representative paradigms, which are genetic algorithm (GA) (Goldberg, 1998), stud genetic algorithm (SGA) (Khatib and Fleming, 1998), differential evolution (DE) (Storn and Price, 1997; Zou et al.,2013; Li and Yin, 2016) earthworm optimization algorithm (EWA) (Wang et al.,2015e) biogeography-based optimization (BBO) (Simon, 2008;Li and Yin, 2012; Saremi et al.,2014; Li and Yin, 2012) and animal migration optimization (AMO) (Li et al.,.2014).

Rhinos are one of the largest mammals in the world. The studies about rhinos have been done in various aspects, involving rhino's space use and the increment of population size. For rhino's space use, synoptic model (Horne et al*., 2008)* is one of the most representative paradigms that is used to estimate its probability of occurrence in a given domain associated with a fixed spatial area (i.e. home range), the spatial distribution of resources, and the occurrence of other animals (Horne et al*.,* 2008) which are called herding density variables (HDVs). With the increment of rhino number, the resources represented by HDVs and owned by each rhino individual are becoming less and less. That is, the fewer recourses, the worse they feel. In our current work, we use rhino comfort index (RCI) to represent this feeling. In other words, RCI is used to measure the goodness of a feasible solution. A good solution is analogous to a rhino with a high RCI, and a poor solution is similar to a rhino with a low RCI.

In this paper, synoptic models and population size updating models are formalized and generalized to a general-purpose metaheuristic algorithm. Accordingly, a new kind of swarm-based algorithm, called Rhino Herd (RH), is proposed for coping with global optimization tasks. Null model in synoptic model is a special kind of model without any influences from others. In RH, null model is considered as the initial herding or the herding before updating. This is followed by rhino modification via synoptic model. Finally, the population size is updated by some population size updating model, and newly-generated rhinos are randomly initialized within the given conditions. The RH is benchmarked by fifteen test optimization problems by comparing it with BBO, and SGA. The results clearly show the superiority of RH in searching for the better function values on most benchmarks over BBO and SGA.

The rest of paper is structured as follows. Section 2 reviews the herding behavior of rhinos in nature, involving synoptic model and population size updating model. Subsequently, Section 3 discusses how the herding behavior of rhinos can be used to formulate a general-purpose metaheuristic algorithm. To fully investigate the performance of RH algorithm, several simulation results comparing the optimal RH algorithm with other optimization methods on fifteen benchmark functions, are presented in Section 4. Finally, Section 5 draws some concluding remarks.

2 Herding behavior of rhinos

Rhinos are one of the biggest mammals in the world, and their weight can reach one ton or more. Rhinos are herbivorous, and they mainly live on in leafy materials. The current rhinos only involve five extant species, and some of the rhinos have two horns, while others have a single horn. Several researchers have done many studies of rhinos from various aspects. Synoptic model of space use and population size updating model are two of the most representative paradigms.

2.1 A synoptic model of space use

To describe rhino space use, a multivariate model, called synoptic model, is proposed to estimate a rhino's probability of occurrence associated with various HDVs, such as a fixed spatial area (i.e., home range), the spatial distribution of resources, and the occurrence of other animals (Horne et al*.,* 2008). In synoptic model, $s(x)$ represents the probability density of finding the rhino at location x during the period of study. For each location, k environmental variables (HDVs) are used as covariates to model a rhino's utilization distribution (Horne et al.,2008). In this model, a null model of space use $f_0(x)$ is applied to describe a rhino's utilization distribution without effects from environmental covariates, which is expressed in the form of an exponential power model (Horne et al.,2008), shown in (1).

$$f_0(x) = \frac{2}{c^2 \pi a^2 \Gamma(c)} \exp\left[-\left(\frac{\| x - \mu \|}{a} \right)^{2/c} \right] \qquad (1)$$

where Γ is the gamma function, μ is the center of the distribution, $a>0$ is the scale parameter, $c>0$ is the shape parameter, and $\|x-\mu\|$ is the distance between x and μ (Horne et al*.,* 2008).

Subsequently, a spatially explicit environmental covariate $H(x)$ is added to the null model $f_0(x)$, where $H(x)$ is defined as a function for describing the environmental covariate. The function $H(x)$ has various forms according to its environmental variables to be described. After introducing one covariate, the synoptic model can be expressed as

$$s(x) = \frac{f_0(x) + \beta H(x) f_0(x)}{\int_x \left[f_0(x) + \beta H(x) f_0(x) \right]} \qquad (2)$$

where β is an estimated selection parameter controlling the magnitude of the effect.

Similarly, after introducing k covariates, the synoptic model of space use can be expressed as

$$s(x)=\frac{f_0(x)\prod_{i=1}^{k}\left(1+\beta_i H_i(x)\right)}{\int_x\left[f_0(x)\prod_{i=1}^{k}\left(1+\beta_i H_i(x)\right)\right]} \quad (3)$$

The denominator of (3) is hard to handle, as it cannot be analytically intractable for most combinations of initial models and environmental covariates (Horne et al., 2008). As an alternative, the landscape can be divided into l discrete grid cells, and the denominator of (3) is therefore calculated as follows:

$$A\sum_{j=1}^{l}\left[f_0(x_j)\prod_{i=1}^{k}\left(1+\beta_i H_i(x_j)\right)\right]\approx\int_x\left[f_0(x)\prod_{i=1}^{k}\left(1+\beta_i H_i(x)\right)\right] \quad (4)$$

where A is the area of each grid cell.

Accordingly, (3) can be approximated as

$$s(x)=\frac{f_0(x)\prod_{i=1}^{k}\left(1+\beta_i H_i(x)\right)}{A\sum_{j=1}^{l}\left[f_0(x_j)\prod_{i=1}^{k}\left(1+\beta_i H_i(x_j)\right)\right]} \quad (5)$$

More information about synoptic model of space use can be found in (Horne et al., 2008).

2.2 Population size updating model

The original data from two sites in South Africa is collected to model the rhino population and predict the rhino number next year (Cromsigt et al., 2002). Population density is thus determined. The predicted rhino number, $n(t)$, and the real population number from the original data, $p(t)$, have the following relationship:

$$p(t)=n(t)+\varepsilon(t) \quad (6)$$

where $\varepsilon(t)$ is an error term between $p(t)$ and $n(t)$.

3 Rhino herd (RH) algorithm

Here, the herding behavior of rhinos described in Section 2, involving synoptic model of space use and population size updating model, is formed to handle optimization problems.

3.1 Synoptic model

In this section, how to use the synoptic model of space use to optimize is given. As aforementioned, the synoptic model is to estimate a rhino's probability of occurrence within a given domain. Here, an updated synoptic model is used to determine the direction of the search for the next iteration. For the jth HDV of rhino i, this updated model can be given as

$$S(X_{i,j})=\frac{f_0(X_i)\left(1+\alpha_i H_i(X_{i,j})\right)}{A\sum_{k=1}^{n}\left[f_0(X_{i,k})\left(1+\beta_k H_k(X_{i,k})\right)\right]} \quad (7)$$

where a_i and β_k are the estimated selection parameter controlling the magnitude of the effect from $H(X)$; n is the population size. Null model $f_0(X)$ and A are defined as above. $H_i(X)$ and $H_k(X)$ are defined as functions for describing the related variables that have influence on rhino i. In our current work, for the sake of simplicity, we set $H_i(X_{i,j})=X_{i,j}$, $H_k(X_{i,k})=X_{i,k}$. That is, $H(X)=X$.

After all the HDVs in rhino i are calculated, the rhino i is updated as

$$X_{i,new}=\begin{cases}X_i+S(X_i)\otimes X_i, & rand>0.5\\ X_i-S(X_i)\otimes X_i, & rand\le 0.5\end{cases} \quad (8)$$

where $X_{i,new}$ is the updated rhino, symbol "$\otimes$ represents pairwise product, and *rand* is a random number drawn from a stochastic distribution.

In order to increase the diversity of the population in the later search, a random term is added to above equation. Therefore, the updated expression can be given as

$$X_{i,new}=\begin{cases}X_i+rand\times S(X_i)\otimes X_i, & rand>0.5\\ X_i-rand\times S(X_i)\otimes X_i, & rand\le 0.5\end{cases} \quad (9)$$

It should be noted that for the center μ in (1), for the jth HDV in μ, it can be calculated as

$$\mu_j=\frac{1}{n}\times\sum_{i=1}^{n}X_{i,j} \quad (10)$$

3.2 Population size updating model

The rhino population number varies each year, and the number is generally becoming larger and larger in nature. This trend can be modeled as above. In this paper, the exponential model is used to update population size, which is given as

$$n_{t+1}=n_t+r\times n_t \quad (11)$$

where r is a constant specific growth rate; n_t and n_{t+1} are population size at generation t and t+1, respectively.

3.3 RH algorithm

Rhino modification operator is a critical operator in RH algorithm, which can loosely be given below.

Algorithm 1 Rhino modification

Begin

Calculate RCI_0 for null model (1).

for i=1 to n (all the rhinos in the herding) **do**

Calculate synoptic model S.

Update the rhino i according to S;

end for i

End

Population size updating is another important operator that updates rhino population size based on a certain rule. The exponential model is provided to calculate the modified population size n' in our current work (Section 3.2).

Algorithm 2 Population size updating

Begin

Calculate n' as per the population size updating model.

for i= n+1 to n' (all the newly-generated rhinos) **do**

Initialize X_i with a randomly generated HDV^m.

end for i

End

According to the analyses above, the schematic framework of RH algorithm can be described as follows.

Algorithm 3 Rhino Herd (RH) Algorithm

Begin

Step 1: Parameters initialization. Firstly, the problem-dependent solutions are mapped to HDVs and rhinos. In addition, an elitism parameter, and the parameters used in null model and population size updating models (see Section 4.1) are initialized.

Step 2: Generate a group of rhinos at random X_0^n Each rhino represents a feasible solution to the problem of interest.

Step 3: Map each rhino to RCI. Each rhino in initial herding is mapped to RCI that can measure the goodness of the rhino.

Step 4: Calculate RCI$_0$. The RCI$_0$ of null model at each generation is calculated. Here, the herding before being modified can be considered as null model at each generation, followed by modifying each rhino based on this null model.

Step 5: Rhino modification. For each rhino in the herding, it is modified by synoptic model (Algorithm 1).

Step 6: Map each rhino to RCI. Each rhino in newly-generated herding is mapped to RCI.

Step 7: Population size updating model. Update population size by using updating model. Rrandomly initialize the newly-generated rhinos and calculate its corresponding RCI for each rhino (see Algorithm 2).

Step 8: Stop or not. Go to **Step 4,** if the termination criterion is not satisfied; terminate the optimization process, if the predefined termination criterion is reached.

End

In Algorithm 3, a rhino comfort index RCI: $X \rightarrow R$ is a measure of goodness of the solution that is represented by the rhino. It should be mentioned that, for most population-based metaheuristic algorithms, RCI is called fitness, and its value is the fitness value. A rhino comfort index of null model RCI$_0$: $N \rightarrow R$ is a measure of goodness of the solution that is represented by the X_0. Here, RCI$_0$ is a special RCI that is different with other RCI. Null model can be formulated in different forms, and there are various ways of calculating RCI$_0$.

4 Simulation results

In this section, the RH is verified by benchmark evaluation in comparison with two methods (BBO (Simon, 2008), and SGA (Khatib and Fleming, 1998) on fifteen test problems (Table 1).

In order to obtain fair results, all the implementations are conducted under the same conditions shown in (Wang et al., 2014a).

The same parameters for RH are set as follows: the area of each grid cell A=1; the constant specific growth rate r=0.04; the scale parameter a=2831; the shape parameter c=0.53; for rhino i, its estimated selection parameter a_i and β_i are set to be its RCI and 1/RCI, respectively. The numbers of generations and initial population size are set to 50 and 50, respectively. In other methods, their parameter settings can be found in (Wang et al., 2014a,b) The dimension is twenty.

Table 1. Benchmark functions.

No.	*Name*	*No.*	*Name*
F01	Ackley	F09	Schwefel 2.26
F02	Alpine	F10	Schwefel 1.2
F03	Griewank	F11	Schwefel 2.22
F04	Holzman 2 function	F12	Schwefel 2.21
F05	Levy 8	F13	Step
F06	Pathological function	F14	Sum function
F07	Perm	F15	Zakharov
F08	Powell		

Metaheuristic algorithms are always based on certain stochastic distribution. Therefore, 50 independent runs are implemented (Table 2). In the following experiments, the best solution is highlighted in **bold**.

From Table 2, RH method has demonstrated its best performance on F01, F05, F08, and F12-F14. At the same time, RH is able to find the best solutions with the smallest Std (standard deviation) on F02, F07, and F11. BBO has shown its best performance on F09-F10. Both of them perform significantly better than SGA. This indicates that RH has a powerful search ability, and can find the fittest solution on most benchmarks.

Table 2. Fitness values obtained by three methods.

Test	BBO	RH	SGA
F01	6.19±0.85	**3.84±0.18**	7.80±1.02
F02	**1.68**±0.85	2.07±**0.60**	3.55±1.70
F03	9.98±4.32	10.59±19.24	**7.91±2.97**
F04	398.30±669.60	7.01E4±5.88E4	**159.80±109.70**
F05	2.64±1.38	**2.40±0.42**	2.52±1.54
F06	5.22±0.55	2.77±0.49	4.98±0.56
F07	6.79E51±2.19E51	6.01E51±**1.33E36**	6.01E51±**1.33E36**
F08	167.00±97.03	**8.82±5.21**	100.10±55.83
F09	**926.40±258.80**	4.52E3±483.70	1.04E3±268.20
F10	**1.16E4±4.00E3**	3.74E4±1.05E4	1.62E4±4.89E3
F11	**3.74**±1.75	4.16±**0.86**	11.94±3.34
F12	48.18±9.23	**1.84±0.26**	43.18±13.20
F13	3.30±1.23	**1.12±0.33**	5.26±1.54
F14	89.80±26.06	**6.88±2.87**	121.20±52.95
F15	**137.70**±43.15	217.50±174.50	224.40±65.28

Moreover, the convergent processes of three most representative algorithms on the most representative benchmarks can be given as follows (Figures 1 and 2).

Figure 1 shows the convergent history of F01-F02, F05, and F08. For F1 case, it can be easily observed that RH11, BBO and SGA rank the first, the second, and the third, respectively. For F02 and F05 cases, although all the three algorithms converge to the similar final solutions, RH has the fastest convergent speed, and it can find the best solution within ten generations. For F08 case, RH has a stable convergence speed, and it can find the final best solution after BBO and SGA have been trapped into the premature status.

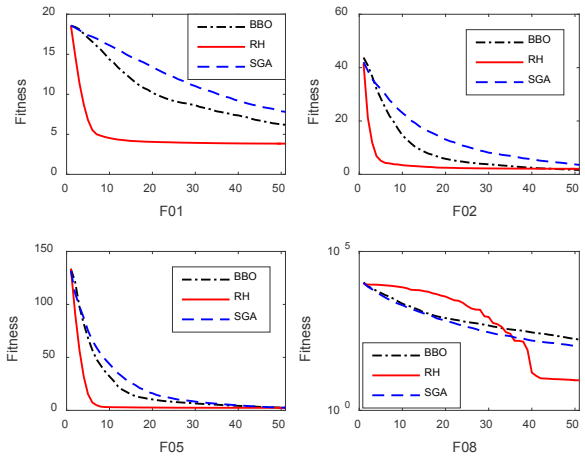


Figure 1. Convergent curves of the benchmarks F01, F02, F05, and F08.

Figure 2 shows the convergent history of benchmarks F11-F14. For F11 and F13 cases, although all the three algorithms can produce similar final solutions, RH has the fastest convergence speed, and it can find the best solution within ten generations. For F12 case, it is clearly visible that RH has a much better solution than BBO and SGA, which have similar optimization performances. For F14 case, RH can eventually find the optimal solution.

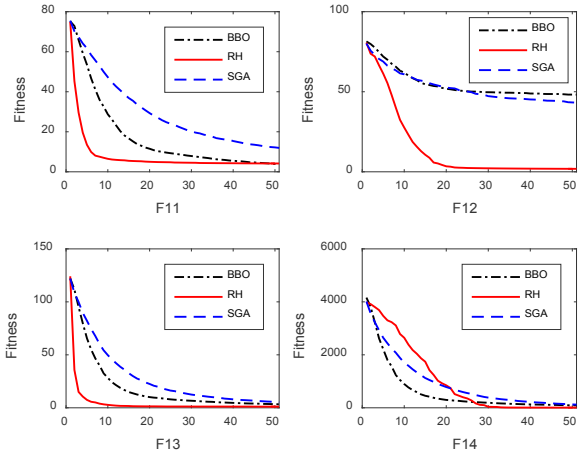


Figure 2. Convergent curves of the benchmarks F11-F14.

5 Discussions and conclusions

We have shown how rhino herding, the research of synoptic model and population size updating model, can be used to develop a novel algorithm for optimization. This new family of algorithms is called RH, which has been benchmarked by fifteen test problems. The results have shown RH's competitive performance in comparison with other two state-of-the-art algorithms. Unfortunately, we cannot conclude RH algorithm is universally better than other two algorithms, or *vice versa*, as per the no free lunch theorem. However, the good performance of RH algorithm in comparison with two other algorithms on fifteen benchmarks demonstrates that it is well capable of addressing practical problems successfully.

In our current work, the influence of some herding density variables is exerted on the null model. Other factors, such as the ratio of male and female, the sun, and landscape will be included in the improved synoptic model as herding density variables.

Another bottleneck of many algorithms is computational requirements. How to reduce computation efforts is highly worthy of in-depth study for an algorithm.

Acknowledgements

This work was supported by Jiangsu Province Science Foundation for Youths (No. BK20150239) and National Natural Science Foundation of China (No. 61503165).

References

J. P. Cromsigt, J. Hearne, I. M. Heitkönig, and H. H. Prins.

Using models in the management of black rhino populations. *Ecological modelling,* 149(1):203-211, 2002.

Z. Cui and X. Gao. Theory and applications of swarm intelligence. *Neural Computing & Applications,* 21(2):205-206, 2012.

M. Dorigo, V. Maniezzo, and A. Colorni. Ant system: optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* 26(1):29-41, 1996.

H. Duan, W. Zhao, G. Wang, and X. Feng. Test-sheet composition using analytic hierarchy process and hybrid metaheuristic algorithm TS/BBO. *Mathematical Problems in Engineering,* 2012:1-22, 2012.

Y. Feng, G.-G. Wang, S. Deb, M. Lu, and X. Zhao. Solving 0-1 knapsack problem by a novel binary monarch butterfly optimization. *Neural Computing and Applications*, 28(7):1619-1634, 2017.

Y.-H. Feng, G.-G. Wang, Q. Feng, and X.-J. Zhao, An effective hybrid cuckoo search algorithm with improved shuffled frog leaping algorithm for 0-1 Knapsack problems. *Computational Intelligence and Neuroscience,* 2014(2014):1-17, 2014.

A. H. Gandomi and A. H. Alavi. Multi-stage genetic programming: a new strategy to nonlinear system modeling. *Information Sciences,* 181(23):5227-5239, 2011.

A. H. Gandomi, X.-S. Yang, and A. H. Alavi. Mixed variable structural optimization using firefly algorithm. *Computers & Structures,* 89(23-24):2325-2336, 2011.

A. H. Gandomi and A. H. Alavi. Krill herd: a new bio-inspired optimization algorithm. *Communications in Nonlinear Science and Numerical Simulation,* 17(12): 4831-4845, 2012.

A. H. Gandomi, S. Talatahari, F. Tadbiri, and A. H. Alavi. Krill herd algorithm for optimum design of truss structures. *International Journal of Bio-Inspired Computation,* 5(5):281-288, 2013.

X. Z. Gao, X. Wang, S. J. Ovaska, and K. Zenger, A hybrid optimization method of harmony search and opposition-based learning. *Engineering Optimization,* 44(8):895-914, 2012a.

X.-Z. Gao, X. Wang, T. Jokinen, S. J. Ovaska, A. Arkkio, and K. Zenger. A hybrid optimization method for wind generator design. *International Journal of Innovative Computing, Information and Control,* 8(6):4347-4373, 2012b.

M. Ghetas, H.-Y. Chan, and P. Sumari, Harmony-based monarch butterfly optimization algorithm. In *Proceedings of the 2015 IEEE International Conference on Control System, Computing and Engineering* (ICCSCE), pages 156-161, November 2015.

D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine learning*. Addison-Wesley, New York, 1998.

L. Guo, G.-G. Wang, H. Wang, and D. Wang. An effective hybrid firefly algorithm with harmony search for global numerical optimization. *The Scientific World Journal,* 2013(2013):1-10, 2013.

L. Guo, G.-G. Wang, A. H. Gandomi, A. H. Alavi, and H. Duan. A new improved krill herd algorithm for global numerical optimization. *Neurocomputing,* 138: 392-402, 2014.

J. S. Horne, E. O. Garton, and J. L. Rachlow. A synoptic model of animal space use: Simultaneous estimation of home range, habitat selection, and inter/intra-specific relationships. *Ecological Modelling,* 214(2-4):338-348, 2008.

D. Karaboga and B. Basturk. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *Journal of Global Optimization,* 39(3):459-471, 2007.

A. Kaveh, R. Sheikholeslami, S. Talatahari, and M. Keshvari-Ilkhichi. Chaotic swarming of particles: a new method for size optimization of truss structures. *Advances in Engineering Software,* 67:136-147, 2014.

J. Kennedy and R. Eberhart, Particle swarm optimization. In *Proceeding of the IEEE International Conference on Neural Networks*, Perth, Australia, pages 1942-1948, 1995.

W. Khatib and P. Fleming. The stud GA: A mini revolution?, *Parallel Problem Solving from Nature - PPSN V*. Lecture Notes in Computer Science. A. Eiben, T. Bäck, M. Schoenauer and H.-P. Schwefel, eds., 683-691, London, UK: Springer Berlin Heidelberg, 1998.

X. Li, and M. Yin. Parameter estimation for chaotic systems by hybrid differential evolution algorithm and artificial bee colony algorithm. *Nonlinear Dynamics,* 77(1-2): 61-71, 2014.

X. Li, and M. Yin. Multiobjective binary biogeography based optimization for feature selection using gene expression data. *IEEE Transactions on NanoBioscience,* 12(4): 343-353, 2013a.

X. Li, and M. Yin. An opposition-based differential evolution algorithm for permutation flow shop scheduling based on diversity measure. *Advances in Engineering Software,* 55: 10-31, 2013b.

X. Li, J. Wang, and M. Yin. Enhancing the performance of cuckoo search algorithm using orthogonal learning method. *Neural Computing and Applications,* 24(6): 1233-1247, 2013.

X. Li, and M. Yin. Modified cuckoo search algorithm with self adaptive parameter method. *Information Sciences,* 298: 80-97, 2015.

Z.-Y. Li, J.-H. Yi, and G.-G. Wang. A new swarm intelligence approach for clustering based on krill herd with elitism strategy. *Algorithms,* 8(4):951-964, 2015.

X. Li and M. Yin. Modified differential evolution with self-adaptive parameters method. *Journal of Combinatorial Optimization,* 31(2):546-576, 2016.

X. Li, and M. Yin. Multi-operator based biogeography based optimization with mutation for global numerical optimization, *Computers & Mathematics with Applications,* 64(9):2833-2844, 2012a.

X.-T. Li, and M.-H. Yin. Parameter estimation for chaotic systems using the cuckoo search algorithm with an orthogonal learning method. *Chinese Physics B,* 21(5): 050507, 2012b.

X. Li, J. Zhang, and M. Yin. Animal migration optimization: an optimization algorithm motivated by elephant herding behavior. *Neural Computing and Applications*, 24(7-8):1867-1877, 2014.

X. Meng, Y. Liu, X. Gao, and H. Zhang. A new bio-inspired algorithm: chicken swarm optimization. *Advances in Swarm Intelligence*, Lecture Notes in Computer Science Y. Tan, Y. Shi and C. C. Coello, eds. 86-94: Springer International Publishing, 2014.

X.-B. Meng, X. Z. Gao, L. Lu, Y. Liu, and H. Zhang. A new bio-inspired optimisation algorithm: bird swarm algorithm. *Journal of Experimental & Theoretical Artificial Intelligence*, 1-15, 2015.

S. Mirjalili, S. M. Mirjalili, and A. Lewis. Let a biogeography-based optimizer train your Multi-Layer Perceptron. *Information Sciences,* 269: 188-209, 2014a.

S. Mirjalili, G.-G. Wang, and L. d. S. Coelho. Binary optimization using hybrid particle swarm optimization and gravitational search algorithm. *Neural Computing and Applications,* 25(6):1423-1435, 2014b.

S. Mirjalili and A. Lewis. S-shaped versus V-shaped transfer functions for binary Particle Swarm Optimization. *Swarm and Evolutionary Computation,* 9:1-14, 2013.

S. Mirjalili. The ant lion optimizer. *Advances in Engineering Software,* 83:80-98, 2015.

S. Mirjalili, S. M. Mirjalili, and X.-S. Yang. Binary bat algorithm. *Neural Computing and Applications,* 25(3-4): 663-681, 2013.

S. Mirjalili, S. M. Mirjalili, and A. Hatamlou. Multi-verse optimizer: a nature-inspired algorithm for global optimization. *Neural Computing and Applications,* 27(2): 495-513, 2016.

S. Mirjalili. Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. *Neural Computing and Applications,* 27(4):1053-1073, 2016.

S. Mirjalili, S. M. Mirjalili, and A. Lewis. Grey wolf optimizer. *Advances in Engineering Software,* 69:46-61, 2014c.

S. Saremi, S. Mirjalili, and A. Lewis. Biogeography-based optimisation with chaos. *Neural Computing and Applications,* 25(5):1077-1097, 2014.

S. Saremi, S. Z. Mirjalili, and S. M. Mirjalili. Evolutionary population dynamics and grey wolf optimizer. *Neural Computing and Applications,* 26(5):1257-1263, 2014.

D. Simon. Biogeography-based optimization. *IEEE Transactions on Evolutionary Computation,* 12(6):702-713, 2008.

R. Storn and K. Price, Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization,* 11(4):341-359, 1997.

G.-G. Wang, L. Guo, H. Duan, L. Liu, and H. Wang. The model and algorithm for the target threat assessment based on Elman_AdaBoost strong predictor. *Acta Electronica Sinica,* 40(5):901-906, 2012a.

G.-G. Wang, H. E. Chu, and S. Mirjalili. Three-dimensional path planning for UCAV using an improved bat algorithm. *Aerospace Science and Technology,* 49:231-238, 2016a.

G. Wang, L. Guo, H. Duan, L. Liu, H. Wang, and J. Wang. A hybrid meta-heuristic DE/CS algorithm for UCAV path planning. *Journal of Information and Computational Science,* 9(16):4811-4818, 2012b.

G. Wang, L. Guo, H. Wang, H. Duan, L. Liu, and J. Li. Incorporating mutation scheme into krill herd algorithm for global numerical optimization. *Neural Computing and Applications,* 24(3-4):853-871, 2014a.

G.-G. Wang, L. Guo, A. H. Gandomi, G.-S. Hao, and H. Wang. Chaotic krill herd algorithm. *Information Sciences,* 274:17-34, 2014b.

G.-G. Wang, A. H. Gandomi, A. H. Alavi, and S. Deb. A hybrid method based on krill herd and quantum-behaved particle swarm optimization. *Neural Computing and Applications,* 27(4):989-1006, 2016a.

G.-G. Wang, A. H. Gandomi, X.-S. Yang, and A. H. Alavi. A novel improved accelerated particle swarm optimization algorithm for global numerical optimization. *Engineering Computations,* 31(7):1198-1220, 2014c.

G.-G. Wang, S. Deb, X.-Z. Gao, and L. d. S. Coelho. A new metaheuristic optimization algorithm motivated by elephant herding behavior. *International Journal of Bio-Inspired Computation*, 8(6):394, 2016b.

G.-G. Wang, A. H. Gandomi, X. Zhao, and H. C. E. Chu. Hybridizing harmony search algorithm with cuckoo search for global numerical optimization. *Soft Computing*, 20(1):273-285, 2016c.

G.-G. Wang, S. Deb, A. H. Gandomi, Z. Zhang, and A. H. Alavi. Chaotic cuckoo search, *Soft Computing*, 20(9):3349-3362, 2016d.

G.-G. Wang, A. H. Gandomi, X.-S. Yang, and A. H. Alavi. A new hybrid method based on krill herd and cuckoo search for global optimization tasks. *International Journal of Bio-Inspired Computation*, 8(5):2016e.

G.-G. Wang, S. Deb, and L. d. S. Coelho. Elephant herding optimization. In *2015 3rd International Symposium on Computational and Business Intelligence (ISCBI 2015),* Bali, Indonesia: 1-5, 2015a.

G.-G. Wang, B. Chang, and Z. Zhang. A multi-swarm bat algorithm for global optimization. In *2015 IEEE Congress on Evolutionary Computation (CEC 2015)*, Sendai, Japan: 480-485, 2015b.

G.-G. Wang, L. Guo, H. Duan, and H. Wang. A new improved firefly algorithm for global numerical optimization. *Journal of Computational and Theoretical Nanoscience,* 11(2):477-485, 2014d.

G.-G. Wang, S. Deb, and Z. Cui. Monarch butterfly optimization. *Neural Computing and Applications*, 2015c.

G.-G. Wang, G.-S. Hao, S. Cheng, and Q. Qin. A discrete monarch butterfly optimization for Chinese TSP problem. *Advances in Swarm Intelligence*, 2016e.

G.-G. Wang, S. Deb, X. Zhao, and Z. Cui. A new monarch butterfly optimization with an improved crossover operator. *Operational Research: An International Journal*, 2016f.

G.-G. Wang, A. H. Gandomi, and A. H. Alavi. A chaotic particle-swarm krill herd algorithm for global numerical optimization. *Kybernetes,* 42(6):962-978, 2013.

G.-G. Wang, A. H. Gandomi, A. H. Alavi, and G.-S. Hao. Hybrid krill herd algorithm with differential evolution for global numerical optimization. *Neural Computing and Applications,*25(2):297-308, 2014e.

G.-G. Wang, A. H. Gandomi, and A. H. Alavi. An effective krill herd algorithm with migration operator in biogeography-based optimization. *Applied Mathematical Modelling,* 38(9-10):2454-2462, 2014f.

G-G Wang, A. H. Gandomi, and A. H. Alavi. Stud krill herd algorithm. *Neurocomputing,* 128: 363-370, 2014g.

G.-G. Wang, A. H. Gandomi, A. H. Alavi, and S. Deb. A multi-stage krill herd algorithm for global numerical optimization. *International Journal on Artificial Intelligence Tools,* 25(2):1550030, 2016g.

G.-G. Wang, S. Deb, A. H. Gandomi, and A. H. Alavi. Opposition-based krill herd algorithm with Cauchy mutation and position clamping. *Neurocomputing,* 177: 147-157, 2016h.

G.-G. Wang, S. Deb, and L. d. S. Coelho. Earthworm

optimization algorithm: a bio-inspired metaheuristic algorithm for global optimization problems. *International Journal of Bio-Inspired Computation*, 2015d.

X. S. Yang. *Nature-inspired metaheuristic algorithms*, 2nd ed., Luniver Press, Frome, 2010.

X. S. Yang, Firefly algorithm, stochastic test functions and design optimisation. *International Journal of Bio-Inspired Computation,*2(2):78-84, 2010.

X. S. Yang and S. Deb, Cuckoo search via Lévy flights. In *Proceeding of World Congress on Nature & Biologically Inspired Computing (NaBIC 2009),* Coimbatore, India:210-214, 2009.

J.-W. Zhang and G.-G. Wang. Image matching using a bat algorithm with mutation. *Applied Mechanics and Materials,* 203(1):88-93, 2012.

X. Zhao, B. Song, P. Huang, Z. Wen, J. Weng, and Y. Fan. An improved discrete immune optimization algorithm based on PSO for QoS-driven web service composition, *Applied Soft Computing,* 12(8):2208-2216, 2012.

X. Zhao. A perturbed particle swarm algorithm for numerical optimization. *Applied Soft Computing,* 10(1):119-124, 2010.

D. Zou, L. Gao, S. Li, and J. Wu. Solving 0-1 knapsack problem by a novel global harmony search algorithm. *Applied Soft Computing,* 11(2):1556-1564, 2011.

Static Stability of Double-Spiral Mobile Robot over Rough Terrain

Naohiko Hanajima[1] Taiki Kaneko[2] Hidekazu Kajiwara[3] Yoshinori Fujihira[1]

[1]College of Design and Manufacturing Technology, Muroran Institute of Technology, Japan,
{hana,yfuji}@mmm.muroran-it.ac.jp
[2]Division of Production Systems Engineering, Muroran Institute of Technology, Japan
[3]College of Information and Systems, Muroran Institute of Technology, Japan,
{kajiwara}@mmm.muroran-it.ac.jp

Abstract

In this paper, we investigate static stability for a double-spiral mobile robot. It is a new locomotion mechanism suitable for the wetlands that suppresses damage to vegetation and does not sink in the mud. The robot walks on the spirals, which play the role of footholds for the mobile robot. To overcome rough terrain locomotion, we need to ensure the stability of the walking motion on the sloping ground. In this study, we applied normalized energy stability margin (NESM) to the double-spiral mobile robot in order to investigate its static stability over rough terrain. The procedure to derived the NESM value were shown from the point of view of the vector calculation. In the numerical case study, we drew NESM maps to investigate the static stabilities when the inclination of the slope varied or the pose and orientation of the robot changed. We adopted a moment in the swing phase where the stability of the robot's balance was easily lost. We found that the robot has sensitive directions in terms of stability. Planning the route and motion of the robot in the rough terrain could help maintain its stability.

Keywords: wetland, field survey, mobile robot, slope, balance

1 Introduction

The problem regarding the reduction of wetlands areas has come to occupy an important position in environmental conservation (Nakamura et al., 2004; Fujita et al., 2009). To investigate the mechanisms of wetlands degradation trends, much effort is put into field surveys (Musgrave and Binley, 2011; Riddell et al., 2010). The survey area where we focus is deeply covered with alder forests or Sasa (veitchii) and the ground surface is formed by a thick pile of withered Sasa stems on the muddy soil and water. It is difficult for researchers to walk around huge areas in the wetlands with large quantities of survey tools. Therefore, a technical support system for field surveys in the wetlands is required.

Recently a double-spiral mobile architecture has been proposed (Hanajima et al., 2009, 2016). It is a new locomotion mechanism suitable for the wetlands that minimizes damage to the vegetation and does not sink in the mud. It consists of two pairs of spirals and one quadruped mobile robot. Each pair of spirals plays the role of footholds for the mobile robot. By traveling at a higher distance from the ground over the spirals, the robot can avoid strong resistance from the dense and hard-stemmed plants. In addition, because the spiral is supported by several contact points on the ground and intermediates between the robot and the muddy ground, the robot never touches the ground and barely sinks.

The quadruped mobile robot has gantry-shaped legs for static walking on the spirals. Each of the legs is connected to a body platform with a rotational joint and two prismatic joints in perpendicular arrangement. Therefore, while all four legs maintain their position, the body platform can change its position and orientation in the horizontal plane while staying level. This provides the body platform with the capacity for arbitrary planar motion in its own plane.

The static walking motion of the double-spiral mobile robot on the flat ground has been performed successfully in numeric simulation. However, rough terrain is inherent in the survey area. We need to account for the stability of the walking motion on rough terrain.

Several stability criteria have been proposed in the field of multi-legged and multi-wheeled systems. For quadruped robots, stability margin was proposed in a very early stage (McGhee and Frank, 1968). It represents the distance between the projection of the center of gravity (COG) on the ground and the border of the supporting polygon. The tip-over stability margin, which is also called force-angle stability margin, is given by product of the magnitude of the net force acting on the COG and the minimum of the angles formed between the direction of the net force and the direction of the tip-over axis normal (Papadopoulos and Rey, 1996). Energy stability margin evaluates the minimum potential energy required to tumble the robot (Messuri and Klein, 1985). Normalized energy stability margin (NESM) is one of the improvements of it (Hirose et al., 2001). It is given by dividing the robot's weight into the energy stability margin for normalization; that is, it becomes the height difference of the COG. In this research, we deal with static stability in walking motion. We do not have to consider the forces other than the gravity acting on the robot. Therefore, we decided to employ NESM for the stability criteria, which

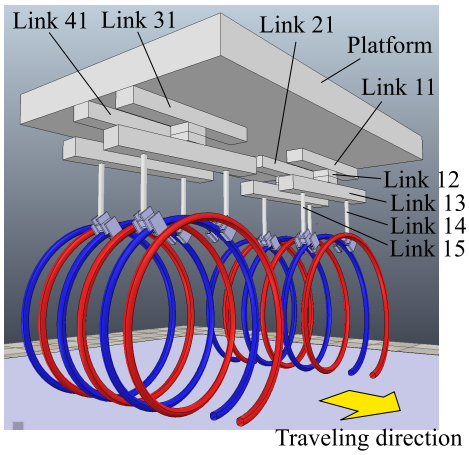


Figure 1. Whole image of the double-spiral robot and definition of links.

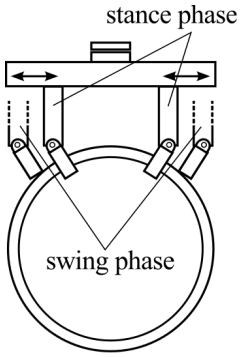


Figure 2. Stance phase and swing phase produced by motion of a gantry-shaped mechanism.

only consider the height of the COG of the robot (Kaneko and Hanajima, 2016).

For a quadruped robot, a swing phase is one of the least stable postures in the walking motion. A creep gait is a well-known static gait in which only one leg can be lifted at any given time (McGhee and Frank, 1968). For a quadruped, when one leg is lifted, the remaining three legs must support the weight of its body. The ground contact points of the three legs form a supporting polygon. Before the swing leg is lifted, the COG of the robot needs to be inside of the support polygon. At the same time, to enhance static stability, it is expected that the COG is located where its NESM is maximized.

In this study, we investigate the NESM values of the double-spiral mobile robot in the swing phase on several gradients of slope from the numerical case study. The knowledge obtained by the case study can be applied to the stable motion planning for the robot's static walking.

In the rest of this paper, first, we review the structure of the double-spiral mobile robot and its kinematics. Next, we introduce the NESM, followed by the numerical case study. Then, we show the results and discuss the potential of the stability of the robot.

2 Structure and Kinematics of the Double-Spiral Mobile Robot

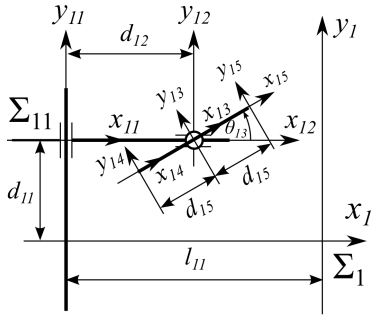
2.1 Structure

Figure 1 shows the structure of the double-spiral mobile robot in the dynamics simulator. As mentioned in the introduction, it consists of two pairs of spirals and one mobile robot. The mobile robot possesses four legs with a gantry-shaped mechanism. Links 11, 21, 31, and 41 move in the direction of travel with respect to the platform with prismatic joints J11, J21, J31, and J41, respectively. Note that the joints are embedded and invisible in Figure 1. We use 2 digits to specify the links and joints of the robot. The first digit denotes a leg number; Leg 1 represents front left; Leg 2, rear left; Leg 3, front right; and Leg 4, rear right. Hereafter, we only explain Leg 1, because every leg has the same structure. Another prismatic joint, J12, is installed at Link 11, perpendicular to the traveling direction. It drives Link 12. The rotational joint, J13, rotates Link 13, which is a part of the legs. As a result, each leg has two prismatic joints orthogonal to each other and one rotational joint for 3 DOF motion in a plane.

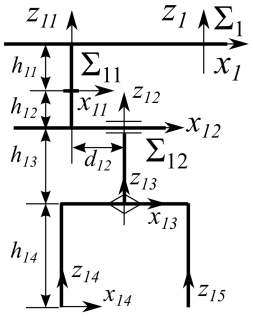
As shown in Figure 1, the end link of each leg has a gantry-shaped mechanism. A pair of vertical links, Links 14 and 15, stands on the spiral in parallel. The distance between them is adjustable by the prismatic joints, J14 and J15. They are driven by one actuator, symmetrically. A gripper is mounted on the lower end of every vertical link. Each gripper requires special mechanisms to grip the rounded rim of the spiral stably so as not to slide down in the stance phase. Once the gripper holds the spiral, the leg needs to maintain its foot position even if the platform moves toward a different posture. In the creep gait, Links 14 and 15 take the opening motion in the swing phase and the closing motion in the stance phase, as shown in Figure 2. In the stance phase, the grippers must hold the spiral tightly.

Note that the working space of the grippers mounted on Link 14 and Link 15 forms the plane parallel to the body platform as well. When the position of the spiral and the working space of the grippers are given, their intersections indicate a pair of points to be gripped. Connecting the pair of points forms a line segment. By aligning Link 13 parallel to the line segment and making the distance between Links 14 and 15 equal to the length of the line segment, each gripper is located at the point to be gripped. In total, the leg mechanism should have enough DOF for the gripper to control a specific position. Since the four legs have the same properties, the position and posture angle of the body platform can be decided within the plane parallel to the grippers' working space independently of any gripper's position.

In total, the robot has enough DOF for its body to move in the desired direction. This function is also important to maintain the balance of the robot. To maintain its balance, the platform of the robot needs to be able to move arbitrarily in the horizontal plane as the NESM value be-



(a) top view



(b) rear view

Figure 3. The coordinate system and joints for one leg of the robot.

comes larger, even though the grippers continue to hold the same places on the spirals.

2.2 Forward kinematics

For the sake of mathematical consideration, we need to define a coordinate system at each link of the robot. Figure 3 shows the coordinate systems of the body platform and Leg 1 from the top view (a) and the rear view (b). We represent the coordinate system of the body as Σ_1 and each coordinate system of Link $\langle a \rangle$ as $\Sigma_{\langle a \rangle}$, respectively. The symbol $O_{\langle a \rangle}$ denotes the origin of $\Sigma_{\langle a \rangle}$, and $x_{\langle a \rangle}$, $y_{\langle a \rangle}$, and $z_{\langle a \rangle}$ denote the x, y, and z axes of $\Sigma_{\langle a \rangle}$, respectively, according to the right-handed coordinate system. We define notation of a position vector of a point p with respect to $\Sigma_{\langle a \rangle}$ as ${}^a\boldsymbol{p}$.

Forward kinematics of a robot is usually represented by homogeneous transformation. The 3×3 block matrix, which consists of the first 3 rows and the first 3 columns of a homogeneous transformation matrix, represents a rotation matrix. The 3×1 vector, which consists of the first 3 rows in the fourth column of a homogeneous transformation matrix, represents a position vector. The position vector can be used to calculate the working space of a link.

With reference to Figure 3, the homogeneous transformation matrices between adjoining coordinate systems for Leg 1 are as follows.

$$ {}^{1}\boldsymbol{H}_{11} = \begin{bmatrix} 1 & 0 & 0 & -l_{11} \\ 0 & 1 & 0 & d_{11} \\ 0 & 0 & 1 & -h_{11} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1} $$

$$ {}^{11}\boldsymbol{H}_{12} = \begin{bmatrix} 1 & 0 & 0 & d_{12} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -h_{12} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2} $$

$$ {}^{12}\boldsymbol{H}_{13} = \begin{bmatrix} \cos\theta_{13} & -\sin\theta_{13} & 0 & 0 \\ \sin\theta_{13} & \cos\theta_{13} & 0 & 0 \\ 0 & 0 & 1 & -h_{13} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3} $$

$$ {}^{13}\boldsymbol{H}_{14} = \begin{bmatrix} 1 & 0 & 0 & -d_{15} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -h_{14} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4} $$

$$ {}^{13}\boldsymbol{H}_{15} = \begin{bmatrix} 1 & 0 & 0 & d_{15} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -h_{14} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5} $$

where ${}^a\boldsymbol{H}_b$ represents a homogeneous transformation matrix from coordinate system $\langle a \rangle$ to coordinate system $\langle b \rangle$. The values of h_{11}, h_{12}, h_{13}, and h_{14} are the offset distances in the z_1 axis direction between the x_1 axis, x_{11} axis, x_{12} axis, x_{13} axis, and x_{14} axis, respectively. The value of l_{11} is the offset distance in the x_1 axis direction between the y_1 axis and y_{11} axis. The variables of d_{11}, d_{12}, and d_{15} are the joint displacements in the direction of the y_1 axis, x_{11} axis, and x_{13} axis, respectively. The variable of θ_{13} is the joint angle around the z_{13} axis.

2.3 Gripper positions

The calculation of NESM requires specifying the positions of the grippers. The position vectors of grippers ${}^1\boldsymbol{p}_{14}$ and ${}^1\boldsymbol{p}_{15}$ with respect to Σ_1 are represented as follows.

$$ {}^1\boldsymbol{p}_{14} = \boldsymbol{I}_{3\times4}\,{}^1\boldsymbol{H}_{11}\,{}^{11}\boldsymbol{H}_{12}\,{}^{12}\boldsymbol{H}_{13}\,{}^{13}\boldsymbol{H}_{14}\boldsymbol{p}_O \tag{6} $$

$$ = \begin{bmatrix} -l_{11} + d_{12} - d_{15}\cos\theta_{13} \\ d_{11} - d_{15}\sin\theta_{13} \\ -h_{11} - h_{12} - h_{13} - h_{14} \end{bmatrix} \tag{7} $$

$$ {}^1\boldsymbol{p}_{15} = \boldsymbol{I}_{3\times4}\,{}^1\boldsymbol{H}_{11}\,{}^{11}\boldsymbol{H}_{12}\,{}^{12}\boldsymbol{H}_{13}\,{}^{13}\boldsymbol{H}_{15}\boldsymbol{p}_O \tag{8} $$

$$ = \begin{bmatrix} -l_{11} + d_{12} + d_{15}\cos\theta_{13} \\ d_{11} + d_{15}\sin\theta_{13} \\ -h_{11} - h_{12} - h_{13} - h_{14} \end{bmatrix} \tag{9} $$

where

$$ \boldsymbol{I}_{3\times4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \boldsymbol{p}_O = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{10} $$

The position vectors of the grippers for the other legs are defined in the same manner.

2.4 COG

The calculation of NESM also requires the COG of the whole robot. The robot body is divided into three portions from the aspect of the COG; the body platform, the first

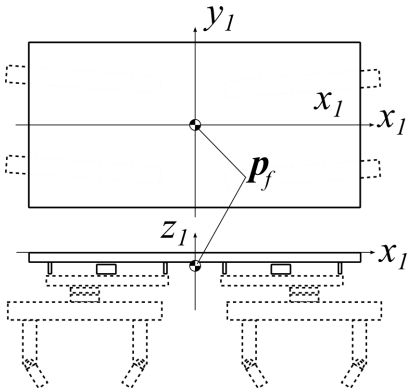


Figure 4. COG of the platform.

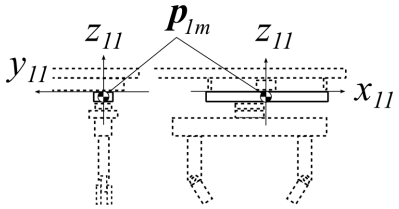


Figure 5. COG of Link 11.

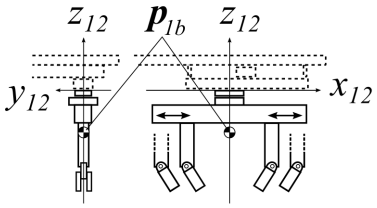


Figure 6. COG of Link 12.

links for the four legs, and the four gantry-shaped legs. We define the position vectors for each portion as $\boldsymbol{p}_f$, $\boldsymbol{p}_{nm}$, and $\boldsymbol{p}_{nb}$ as shown in Figure 4, Figure 5, and Figure 6, respectively, where n denotes the leg number. The position vector of the total COG of the whole robot, $\boldsymbol{p}_c$, is obtained by the following equation.

$$\boldsymbol{p}_c = \frac{m_f \boldsymbol{p}_f + \sum_{n=1}^{4}(m_{nm}\boldsymbol{p}_{nm} + m_{nb}\boldsymbol{p}_{nb})}{m_f + \sum_{n=1}^{4}(m_{nm} + m_{nb})} \qquad (11)$$

where m_f, m_{nm}, and m_{nb} denote the mass of the three portions: the body platform, the first links for the four legs, and the four gantry-shaped legs, respectively.

3 NESM

In this section, we apply the NESM to the double-spiral mobile robot on a slope and show the calculation method.

First, we define a reference coordinate frame Σ_R in the field. It is supposed that the double-spiral mobile robot is located on the slope as shown in Figure 7. The axes x_R, y_R, and z_R denote the principal axes of Σ_R. The positive direction of z_R is upward to the vertical line. The plane consisting of x_R and y_R is set arbitrarily.

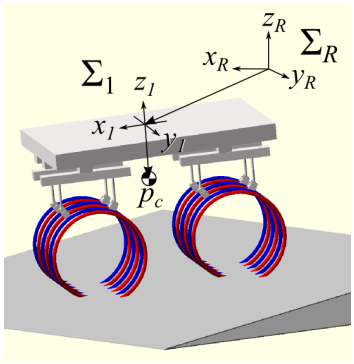


Figure 7. Reference frame and relationship of frames.

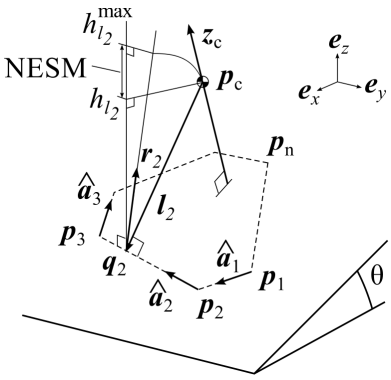


Figure 8. Illustration of normalized energy stability margin.

To derive NESM, the height information in the reference coordinate frame is mandatory. We define the homogenous transfer matrix ${}^R\boldsymbol{H}_1$ from Σ_R to Σ_1, which represents the position and orientation of the whole robot. We can obtain the position vectors with respect to Σ_R by premultiplying the position vector with respect to Σ_1 by ${}^R\boldsymbol{H}_1$. After the position vectors of the COG and the grippers of the four legs are obtained, their coordinates are converted to those with respect to Σ_R using ${}^R\boldsymbol{H}_1$. Hereafter, we use the notation $\boldsymbol{p}_c$ for the position vector with respect to Σ_R.

In the stance phase, all eight grippers hold the spirals. In the swing phase, one pair of grippers releases one of the spirals. Before the moment of release, the body platform needs to be located at the position where its NESM in the swing phase is maximized. The positions of the six grippers that will hold the spiral in the swing phase form a polygon. However, the polygon does not always form a convex polygon. By applying a quickhull algorithm (Barber et al., 1996) to the positions of the six grippers, we can obtain the vertices of a convex polygon that forms a support polygon in the swing phase at the same time. We refer to the vertices of the support polygon as p_i, where i is an index number.

Figure 8 illustrates NESM. Each $\boldsymbol{p}_i$ denotes a position vector relative to the vertex p_i. The support polygon is represented by broken lines. It is important that the index i is ordered clockwise in the top view. When a tip over occurs, one edge of the polygon can be a rotation axis. A

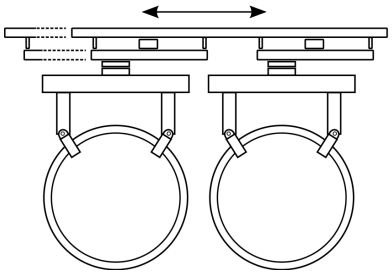


Figure 9. Motion of the platform: front view.

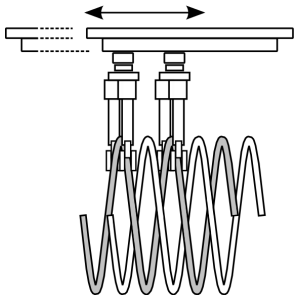


Figure 10. Motion of the platform: side view.

tip-over vector $\boldsymbol{a}_i$ is defined as the following equation so that it is aligned with the edge.

$$\boldsymbol{a}_i = \boldsymbol{p}_{i+1} - \boldsymbol{p}_i \qquad (i = 1, \cdots, n) \tag{12}$$

where $\boldsymbol{p}_{n+1} \equiv \boldsymbol{p}_1$. A tip-over unit vector $\hat{\boldsymbol{a}}_i$ has same direction as $\boldsymbol{a}_i$; that is, $\hat{\boldsymbol{a}}_i = \boldsymbol{a}_i/|\boldsymbol{a}_i|$. A vector $\boldsymbol{z}_c$ is normal to the support polygon passing through the COG of the whole robot. The position vector of the COG is shown as $\boldsymbol{p}_c$. A tip-over vector normal $\boldsymbol{l}_i$ is chosen so that it intersects the COG and the edge regarding $\boldsymbol{a}_i$. It is given by the following equation.

$$\boldsymbol{l}_i = (\boldsymbol{I} - \hat{\boldsymbol{a}}_i \hat{\boldsymbol{a}}_i^T)(\boldsymbol{p}_{i+1} - \boldsymbol{p}_c) \tag{13}$$

where $\boldsymbol{I}$ is the identity matrix. The position vector $\boldsymbol{q}_i$ is chosen where an intersection of $\boldsymbol{l}_i$ and $\boldsymbol{a}_i$. It is obtained by the following equation.

$$\boldsymbol{q}_i = \boldsymbol{l}_i + \boldsymbol{p}_c \tag{14}$$

Unit vectors $\boldsymbol{e}_x$, $\boldsymbol{e}_y$, and $\boldsymbol{e}_z$ are alined to the respective axes of the reference frame. The height of COG with respect to the reference frame, h_{l_i}, is equivalent to the z axis component of $\boldsymbol{p}_c$. Therefore, it is obtained by the following equation.

$$h_{l_i} = \boldsymbol{e}_z^T \boldsymbol{p}_c \tag{15}$$

When the robot tumbles about the axis regarding $\boldsymbol{a}_i$, the locus of COG follows a circular path. The highest point of the locus is just above the axis regarding $\boldsymbol{a}_i$. The vector $\boldsymbol{r}_i$ is oriented from the position regarding $\boldsymbol{q}_i$ to the highest point. A vector given by a cross product of $\boldsymbol{e}_z$ and $\boldsymbol{a}_i$ is the normal vector of the plane including vector $\boldsymbol{r}_i$, which is perpendicular to $\boldsymbol{a}_i$. Therefore, the following relationship is satisfied.

$$\boldsymbol{r}_i = \boldsymbol{a}_i \times (\boldsymbol{e}_i \times \boldsymbol{a}_i) \tag{16}$$

The radius of the circle drawn by the COG is equal to $|\boldsymbol{l}_i|$. Then, the height of the highest point with respect to the reference frame, $h_{l_i}^{\max}$, is obtained by the following equation.

$$h_{l_i}^{\max} = \boldsymbol{e}_z^T (\boldsymbol{q}_i + |\boldsymbol{l}_i| \hat{\boldsymbol{r}}_i) \tag{17}$$

Finally, NESM for all edges of the support polygon is obtained by the following equation.

$$\mathrm{NESM} = \min_i (h_{l_i}^{\max} - h_{l_i}) \sigma_i \tag{18}$$

$$= \min_i \boldsymbol{e}_z^T (\boldsymbol{l}_i + |\boldsymbol{l}_i| \hat{\boldsymbol{r}}_i) \sigma_i \tag{19}$$

where $\hat{\boldsymbol{r}}_i$ is a unit vector in the same manner as $\hat{\boldsymbol{a}}_i$, and σ_i represents a sign of NESM, which is positive when the COG is in the stable side and negative otherwise. σ_i is defined as the following formula.

$$\sigma_i = (\hat{\boldsymbol{r}}_i \times \hat{\boldsymbol{l}}_i)^T \hat{\boldsymbol{a}}_i \tag{20}$$

where $\hat{\boldsymbol{l}}_i$ is a unit vector in the same manner as $\hat{\boldsymbol{a}}_i$.

4 Numerical Case Study

In this section, we investigate the NESM values of the double-spiral mobile robot in the swing phase on several gradients of slope.

4.1 Conditions and methods

Table 1 shows the values used in the numerical case study, such as link parameters, platform sizes, weight, joint variables, and so on. The number n in the subscript denotes the leg number.

The parameters h_{pg}, h_{ln}, and h_{lg} represent positions of the COGs of the three portions. We suppose that the density of each portion is uniform. Therefore, we can assume that $\boldsymbol{p}_f$ and $\boldsymbol{p}_{nm}$ locate on the z_1 axis and z_{n1} axis, respectively, and that $\boldsymbol{p}_{nb}$ locates on the z_{n2} axis because Link n4 and Link n5 move symmetrically. As a result, we have the following:

$${}^1\boldsymbol{p}_f = \begin{bmatrix} 0 \\ 0 \\ h_{pg} \end{bmatrix}, {}^{n1}\boldsymbol{p}_{nm} = \begin{bmatrix} 0 \\ 0 \\ h_{ln} \end{bmatrix}, {}^{n2}\boldsymbol{p}_{nb} = \begin{bmatrix} 0 \\ 0 \\ h_{lg} \end{bmatrix}. \tag{21}$$

In this numerical case study, we set the initial values of the joint parameters as shown in Table 1, such that the double-spiral mobile robot is in the swing phase —i.e., Leg 1 is a swing leg. After that, we move the platform to the position that is taken at each of twenty points at regular intervals within the full stroke of each joint in the x_1 and y_1 directions, as in Figure 9 and Figure 10, respectively. During the movement, the leg positions remain stationary. The ranges of the joint parameters are designated as strokes in Table 1. After calculating the values of NESM

Table 1. The values used in the numerical case study. n in the subscript denotes the leg number.

Item	*Value*	*Item*	*Value*
l_{n1}	0.5[m]	width of platform	1.0[m]
h_{n1}	0.1[m]	depth of platform	0.5[m]
h_{n2}	0.05[m]	height of platform	0.1[m]
h_{n3}	0.1[m]	stroke of joint 1, 2, 4	0.4[m]
h_{n4}	0.5[m]	stroke of joint 3	0.4π[rad]
h_{pg}	0.1[m]	d_{11},d_{31}	0.2[m]
h_{ln}	0.025[m]	d_{21},d_{41}	-0.2[m]
h_{lg}	0.3[m]	d_{n2}	0.0[m]
m_f	100[kg]	θ_{n3}	$\pi/18$[rad]
m_{nm}	5[kg]	d_{14}	0.3[m]
m_{nb}	20[kg]	d_{24},d_{34},d_{44}	0.2[m]

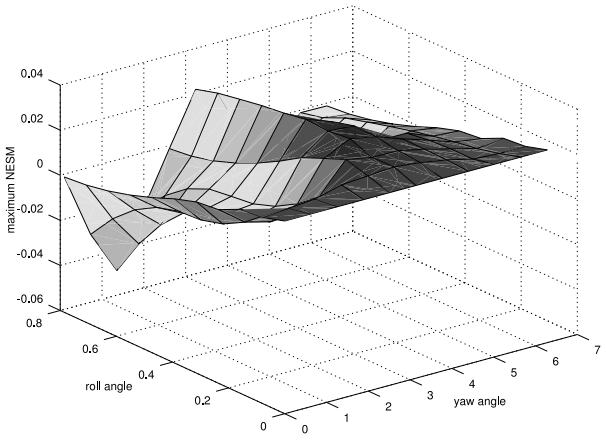


Figure 11. Maximum value map of NESM with respect to the yaw and roll angle of slope gradient.

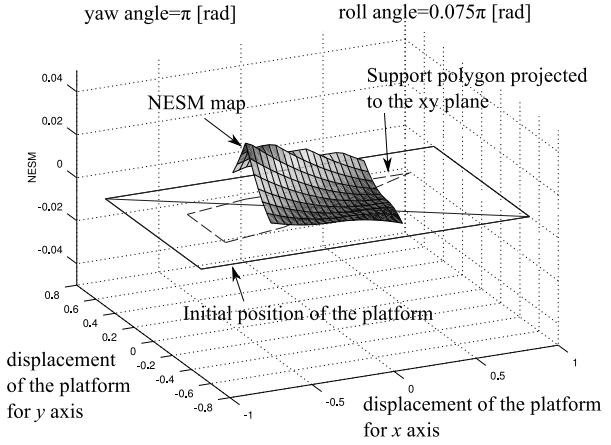


Figure 12. NESM map with respect to the x-y displacement of the platform when the yaw angle is π [rad] and the roll angle is $\frac{3\pi}{40}$[rad] at the largest maximum NESM as shown in Figure 11.

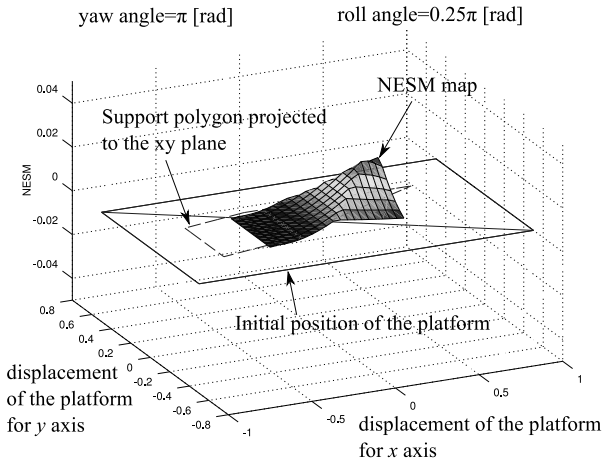


Figure 13. NESM map with respect to the x-y displacement of the platform when the yaw angle is π [rad] and the roll angle is $\frac{\pi}{4}$[rad]. The robot keeps the same direction as Figure 12 —i.e. the yaw angle of π[rad], where the maximum NEMS value is minimized at the roll angle of $\frac{\pi}{4}$[rad].

at each position, the maximum value of NESM and the corresponding position of the platform are obtained.

We prepare the slope conditions by changing two orientation angles of the robot: the yaw angle around z_R and the roll angle around x_R in Σ_R. The roll angle represents the inclination of the slope. The yaw angle represents the front direction of the robot. We set the range of the yaw angle from 0[rad] to 2π[rad] every 0.2π[rad], and the roll angle from 0[rad] to $\frac{\pi}{4}$[rad] every $\frac{\pi}{40}$[rad]. Then we obtain a maximum values map of NESM with respect to the yaw and roll angles.

4.2 Results

Figure 11 shows the maximum values map of NESM with respect to the yaw and roll angles. The largest maximum NESM is observed as 0.0328 when the yaw angle is π[rad] and the roll angle is $\frac{3\pi}{40}$[rad]. When the robot keeps the same direction as the yaw angle of π[rad] and the inclination of the slope changes, the smallest maximum NESM is observed as 0.0232 when the roll angle is $\frac{\pi}{4}$[rad].

Figure 12 shows the NESM map with respect to the x-y displacement of the platform when the yaw angle is π[rad] and the roll angle is $\frac{3\pi}{40}$[rad]. Figure 13 is different from Figure 12 only in that the roll angle is $\frac{\pi}{4}$[rad]. For the sake of comprehension, the projection of the support polygon to the *xy* plane is drawn by a broken line. The solid rectangle represents the initial position of the platform. It is drawn on the zero-level plane of the NESM.

Figure 14 and Figure 15 show wire-frame models of the robot in the cases of maximum NESM in Figure 12 and Figure 13, respectively. Each of the platforms is painted gray. Under the platform, the gantry-shaped leg is represented by solid lines. The gripper positions of Legs 2, 3, and 4 form the support polygon designated by the broken lines.

Figure 16 shows the support polygons, the position of the COGs, and maximum height regarding the NESM re-

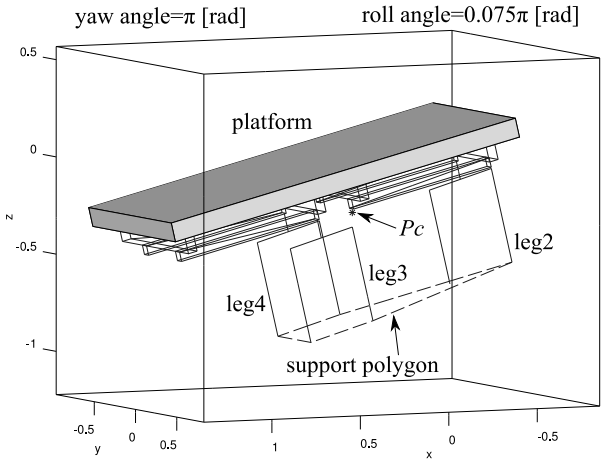


Figure 14. Relationship between COG and support polygon at the maximum NESM in the Figure 12. p_c designates the position of COG.

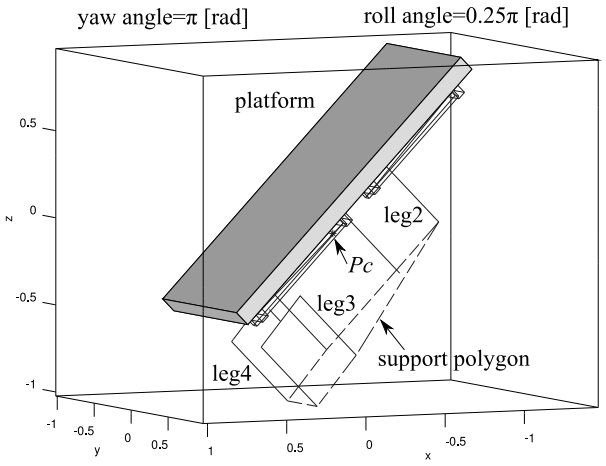


Figure 15. Relationship between COG and the support polygon at the maximum NESM in the Figure 13. p_c designates the position of COG.

lationship in Figure 12 and Figure 13.

5 Discussion

We can observe in Figure 11 that the maximum NESM tends to decrease in general as the roll angle increases — that is, the inclination of the slope becomes larger.

As shown in the preceding figures, the shape of the support polygon of the robot is long and narrow. It is easy to find that a stability margin against tumbling tends to be wider in the longer direction and narrower in the shorter direction. Figure 11 shows this tendency; that is, the maximum NESM is relatively high at the yaw angle around 0[rad] or π[rad] and low around $\frac{\pi}{2}$[rad] or $\frac{3\pi}{2}$[rad]. Especially in the latter case, some of the maximum NESMs become negative when the roll angle is larger than $\frac{\pi}{8}$[rad].

Comparing Figure 12 and Figure 13, the position of the peek in the NESM map is different. According to Figure 16, the tip-over edge in the case of Figure 12 is a swing leg side on the support polygon. It likely happens in the

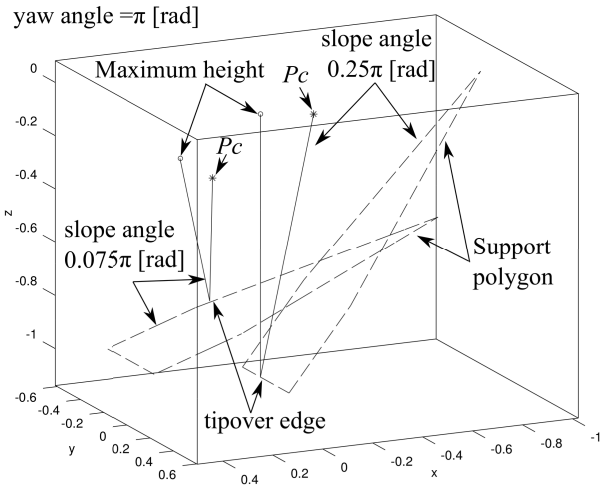


Figure 16. Relationship between COG and the support polygon when NESM is maximum and minimum. p_c designates the position of COG.

swing phase of the walking motion on even terrain. On the other hand, the tip-over edge in the case of Figure 13 is a downhill side of the slope on the support polygon. This means that the tip-over edge of the support polygon depends on the steepness of the slope. To maintain a high enough NESM value during the walking motion, it is important to measure the steepness of the slope and check NESM at every moment.

In a real situation, the robot should move to avoid very steep terrain. The strength of a gantry-shape leg or the maximum force and torque of the actuators must not exceed the design specification. However, as shown in Figure 15, very steep slope requires greater strength or force/-torque. Theoretically, the positive value of NESM can maintain the robot's balance; however, too much inclination may damage the machine.

6 Conclusions

In this paper, we investigated the NESM values of a double-spiral mobile robot in the swing phase on several gradients of the slope from the numerical case study. We derived the support polygon from the kinematics analysis and showed the position vector of COG of the whole robot. Then, we described the way to obtain NESM values from the vector calculation.

The numerical case study revealed that the stability margin against tumbling tended to be wider in the longer direction of the body platform and narrower in the shorter direction. we reported the results and discussed the potential stability of the robot.

Even if the direction of the inclination of the slope was the same, if the degree of the inclination changed, the shape of the NESM map with respect to the position of the platform was also different. The double-spiral mobile robot was able to maintain its foot position while the position of the platform changed. Therefore, it is important to measure the steepness of the slope, to check NESM at ev-

ery moment, and to move the platform in the appropriate direction to maintain a high enough NESM value in the walking motion.

The numerical case study discussed in this paper just showed the NESM map for a specific pose of the robot in a stationary condition. To investigate dynamic stability or to perform the dynamical simulation is left for future work.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 24560282 and 16K06171. The authors also express their sincere gratitude to Center of Environmental Science and Disaster Mitigation for Advanced Research (CEDAR) at Muroran Institute of Technology for meaningful assistance.

References

C. B. Barber, D. P. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Trans. on Mathematical Software*, 22(4):469–483, 1996.

H. Fujita, Y. Igarashi, S. Hotes, M. Takada, T. Inoue, and M. Kaneko. An inventory of the mires of Hokkaido, Japan — their development, classification, decline, and conservation. *Plant Ecology*, 200(1):9–36, 2009.

N. Hanajima, Y. Hayasaka, N. Azumi, K. Kawauchi, M. Yamashita, and H. Hikita. Double spiral propulsion mechanism in wetlands. In *Joint Seminar on Environmental Science and Disaster Mitigation Research 2009*, pages 77–78, 2009.

N. Hanajima, Q. Liu, and H. Kajiwara. A four-legged mobile robot with prismatic joints on spiral footholds. Technical Report 65, Memoirs of the Muroran Inst. of Tech., 2016.

S. Hirose, H. Tsukagoshi, and K. Yoneda. Normalized energy stability margin and its contour of walking vehicles on rough terrain. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA'2001)*, pages 181–186, 2001.

T. Kaneko and N. Hanajima. Evaluation of walking motion on the slope ground using normalized energy stability margin for a double spiral mobile robot. In *Proc. 48th SICE Hokkaido Branch Conf.*, pages 31–32, 2016. (in Japanese).

R. B. McGhee and A. A. Frank. On the stability properties of quadruped creeping gaits. *Mathematical Biosciences*, 3(2): 331–351, 1968.

D. A. Messuri and C. A. Klein. Automatic body regulation for maintaining stability of a legged vehicle during rough-terrain locomotion. *IEEE Trans. Robot. Autom.*, RA-1(3):132–141, 1985.

H. Musgrave and A. Binley. Revealing the temporal dynamics of subsurface temperature in a wetland using time-lapse geophysics. *Journal of Hydrology*, 396(3-4):258–266, 2011.

F. Nakamura, S. Kameyama, and S. Mizugaki. Rapid shrinkage of kushiro mire, the largest mire in japan, due to increased sedimentation associated with land-use development in the catchmen. *Catena*, 55(2):213–229, 2004.

E. G. Papadopoulos and D. A. Rey. A new measure of tipover stability for mobile manipulators. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA'96)*, pages 3113–3116, 1996.

E. S. Riddell, S. A. Lorentz, and D. C. Kotze. A geophysical analysis of hydro-geomorphic controls within a headwater wetland in a granitic landscape, through eri and ip. *Hydrology and Earth System Sciences*, 14(8):1697–1713, 2010.

New Approach based on Simplification and partially fixing of Problem to solve Large Scale Vehicle Routing Problem

Shinya Watanabe[1] Tetsuya Sato[2] Kazutoshi Sakakibara[3]

[1]College of Information and Systems,Muroran Institute of Technology, Japan, sin@csse.muroran-it.ac.jp
[2]Mizuho Information & Research Institute, Inc., Japan,tetsu.en.sato@gmail.com
[3]Department of Information Systems Engineering,Toyama Prefectural University, Japan,
sakakibara@pu-toyama.ac.jp

Abstract

This paper presents a specialized evolutionary approach for large scale vehicle routing problems (VRPs). Our approach includes two original mechanisms; simplification of problem and partially fixing of customers' sequence. The first one tries to simplify the problem by integrating some neighbor customers into one group recursively and to iterate to restore the simplified problem to original one gradually. And second mechanism is to reduce the search space of the problem by fixing a part of customers' sequence. Our approach is designed for an effective search in large scale VRPs by the interaction of these mechanisms. Through applying the proposed approach to some test problems having different characteristics, the effectiveness of our approach is determined in comparison with normal approach (without our these original mechanisms).

Keywords: vehicle routing problem, large scale problem, evolutionary multi-criterion optimization

1 Introduction

Vehicle Routing Problems (VRPs) are well known as NP-hard combinatorial optimization problems arising in many distribution and transportation systems, such as postal delivery, school bus routing, newspaper distribution, etc. VRPs have attracted a great deal of attention since 1970's due to their wide applicability and economic importance(Braysy and Gendreau, 2005; Watanabe and Sakakibara, 2007).

Although the objective of most VRPs' application is to minimize the total area distance, VRPs inherently have multi-objective aspects such as the number of vehicles or the degree of dispersion between the distances of each vehicle. Therefore, there have been many studies using evolutionary multi-criterion optimization (EMO) algorithm to optimize multi-objective VRPs (Watanabe and Sakakibara, 2007; Jozefowiez et al., 2002).

Recently, data size and problem size has become larger scale according to technical advantages of storage performance and cloud technology. Since this trend causes new formidable issues such as the combinatorial explosion and the increment in computational cost, previous approaches are difficult to obtain solutions to fill required quality in real time.

In this research, we propose a new approach dedicated to very large scale VRPs. The proposed approach has two distinguishing mechanisms; simplification of problem and partially fixing of customers' sequence. The first mechanism tries to obtain high quality candidate solutions at early stage by mixing simplification and gradual restoration techniques. In particular, simplification is used firstly to reduce the number of customers apparently by gathering some neighborhood customers to one virtual customer.

Also, the second fixing mechanism is used to reduce search space by fixing a part of customers' sequence and perform an efficient refinement of the obtained solution by applying the first mechanism. The fixing of customers' sequence is very reasonable way to reduce a number of combination in large scale problem.

Through some test examples of Cordeau's instances from VRP website[1] , we showed that the proposed approach can obtain high quality solution in the case of very large problem size.

2 Vehicle Routing Problem

There are many different types in vehicle routing problems (VRPs) based on the type of handling constraints. This paper deals with the multiple depots vehicle routing problems (MDVRPs) (Luo et al., 2013) having only a capacity constrain, which is generally called the capacitated VRPs (CVRPs).

The definitions of CVRPs and MDVRPs can be described as below.

2.1 Capacitated Vehicle Routing Problems

There are many kinds of VRPs according to the type of constraints. CVRPs having only a capacity constrain can be defined as follows (Braysy and Gendreau, 2005):

- All vehicles start from the depot and visit the assigned customer points, then return to the depot. Here, a route is formed by the sequence of the depot

[1]VRPwebsite http://www.bernabe.dorronsoro.es/vrp/.

and the customer points visited by a vehicle. Therefore the number of vehicles is same as the number of route. Moreover, each customer is visited only once by exactly one vehicle.

- Each customer asks for a weight $w_i (i = 1, \ldots, N)$[2] of goods and a vehicle of capacity W is available to deliver the goods. In this paper, we used the same capacity W for all vehicles.
- A solution of the CVRPs is a collection of routes where the total route demand is at most W.

In this paper, we treated CVRPs having multiple depots (MDVRPs). In MDVRPs, departure and arrival points of each vehicle should be same.

VRPs have a number of objectives, such as minimization of the total travel distance, minimization of the number of routes, minimization of the duration of the routes, etc. In this paper, we treated VRPs as two objective problem; minimization of the total travel distance (F_{dist}) and the variance of travel distance of each vehicle F_{var}. The formulas of these objectives are as follows:

$$\text{minimize } F_{\text{dist}} = \sum_{m=1}^{M} c^m \quad (1)$$

$$\text{minimize } F_{\text{var}} = \frac{1}{M} \sum_{m=1}^{M} (\bar{c} - c^m)^2 \quad (2)$$

where M is the total number of routes and c^m and $\bar{c}$ indicate m th route distance and the average value of c^m, respectively.

2.2 Multi-Depot Vehicle Routing Problems

MDVRPs add one more constraint to CVRPs relative to multi-depots. This constraint is that starting point and ending point of each vehicle should be the same depot.

3 Proposed approach

In this paper, we proposed a new approach specialized for large scale VPRs. The main feature of the proposed approach is to have two distinguishing mechanisms; simplification and fixing mechanisms. The first one is used for obtaining high quality solutions at early search phase and the second one is assumed to perform an efficient refinement of solutions at late search phase.

In this section, the details of the proposed approach were explained.

3.1 The flow of the proposed approach

Here, we described the detail of our approach as follows.

Step 0: The setting of initial parameters

- N The number of initial population
- G_S The upper period of the search stagnation (the timing parameter of dissolving cluster)

[2] N is the number of customers.

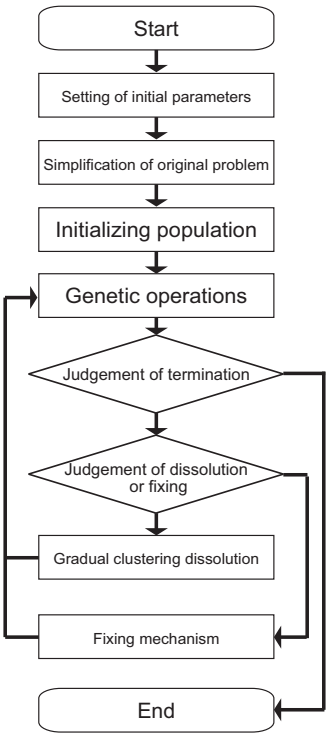


Figure 1. Flowchart of the proposed approach.

- G_R The upper period of the search stagnation (the timing parameter of fixing mechnism)
- E_{Final} Terminal criteria

Step 1: Simplification of the problem

In order to simplify the original problem, neighborhood customers would be gathered to one virtual customer by clustering method and this gathering process would be repeated until the number of customer reaches the pre-defined number.

Step 2: Initializing population

Create N initial individuals and set the parameter t representing the generation $t = 0$ and the parameter g_s representing the term of search stagnation $g_s = 0$.

Step 3: Updating solutions

Solutions would be improved by genetic operations. And the solution having the best value of $F_{\text{sum}^{\text{T}}}$ at T generation would be found. If $F_{\text{best}^{\text{T}-1}}$ and F^T_{best} are same, the value of g_s is incremented by one ($g_s = g_s + 1$) and in other case ($F^{T-1}_{\text{best}} \neq F^T_{\text{best}}$), reset the value of $g_s (g_s = 0)$. However g_s would be kept 0 until this step repeat G_r times.

Step 4: Judgment of termination

If termination condition is satisfied, terminate this optimization process. Also, if $G_S \leq g_s$ and $p \neq 1$, go to Step 5 (clustering dissolution phase), and in the case of $G_R \leq g_s$, $p = 1$, go to Step 6 (fixing mechinsm). Otherwise, go back to Step3.

Step 5: Cluster dissolution

In order to reinstate simplified problem, gradual dis-

solution mechanism would dissolve clustering in a limited area (the details of this mechanism is described below). After resetting $g_s = 0$, go back to Step4.

Step 6: Fixing and unfixing of customers' sequence

In order to reduce search space by fixing a part of customers' sequence, fixing mechanism would implement the fixing or unfixing of customers' sequence(the details of this mechanism is described below). After resetting $g_s = 0$, go back to Step4.

Hereinafter, simplification of problem of of Step1, gradual cluster dissolution mechanism of Step 5 and fixing mechanism of Step 6 are described in detail.

3.2 Simplification of the problem

Our simplification technique tries to reduce the number of customers apparently by gathering some neighborhood customers to one virtual customer. In this simplification, clustering method is used to group a set of customers and some unique features in VRPs are considered.

As a results of analyzing best known solutions of some different famous CVRP instances, we could find the relation between depots and customers. The customers near depot tend to belong different routes even though they are very close. On the other hand, in the case of the customers being far from depot, customers in the near distance tend to belong same route.

Therefore we designed our simplification that the customers being father away from depots result in a higher rate of grouping customers (creating sets).

Also, in this simplification, we restrict grouping customers in case that their distance is over a certain amount and the total weight of grouping customers is over a weight capacity of vehicle.

The flow of this simplification is shown as below.

Step 0: Creating initial clusters

Creating initial clusters $C_i (i = 1, \ldots, N)$ having each one customer and setting the initial clusters' total load $w^{C_i} (i = 1, \cdots, N)$. Here, set the number of customer in each cluster $P_i = 1$ and current total number of clusters $N^C = N$.

Step 1: Calculating the maximum number of customers for each cluster

Calculating the maximum number of customers $P_i^{\max}$ for each cluster. This $P_i^{\max}$ is defined according to the distance between cluster and the depot. We set larger $P_i^{\max}$ of cluster larger distance from depot.

Step 2: Selecting the target cluster

Selecting the target cluster satisfying below equation (1) from unselected clusters.

$$\max_{i=1,\cdots,N^C} d_{(C_i,C_0)} \tag{3}$$

where C_0 is depot and $d_{(C_i,C_j)}$ indicates the shortest distance between two clusters. If there are no clusters which satisfy the above equation, finish this procedure.

Step 3: Judgment of termination

If $N^C \leq N^P$, finish this procedure. Otherwise, go to next step.

Step 4: Finding set of clusters to combine

Selecting set of clusters to combine. Here, C_i and C_j satisfying below equation are selected.

$$\begin{aligned} D^T \quad & \geq \min_{j=1,\cdots,N^C | j \neq i} d_{(C_i,C_j)} \\ s.t. \quad & W \geq w^{C_i} + w^{C_j} \\ & (j = 1, \cdots, N^C \mid i \neq j) \\ & P_i^{\max} \geq P_i + P_j \\ & (j = 1, \cdots, N^C \mid i \neq j) \end{aligned} \tag{4}$$

where D^T is the upper limit distance of combing clusters. Also, W indicates the upper load capacity and $d_{(C_i,C_j)}$ represents the distance between two clusters.

Step 5: Combination of two clusters

Combining two clusters (C_i and C_j) and updating cluster's total load and the number of customer within cluster. Specifically, the information of C_j is integrated into that of C_i and remove C_j. After renewing $N^C = N^C - 1$, go back to Step 3.

As mentioned above, this simplification tries to combine two adjacent clusters into one until the total number of clusters reaches the predefined number.

3.3 Gradual cluster dissolution

This technique gradually restore simplified problem to original one. If the simplified problem is restored at once, the differences between before and after problems is so large that the solutions of prior problem wouldn't contribute to the search in a new integrated problem. Therefore, we try to control the differences between before and after problems small by gradually restoration. In the case that the difference of two problems is small, the solutions of prior problem would be good seed solutions in antecedent one and this would tend toward an effective search.

There two key points in this dissolution; the timing of dissolving cluster and the determination which cluster is dissolved next. For former point, we use the term of search stagnation as the timing of dissolving cluster. Specific conditions of dissolution timing are that a best incumbent solution having best F_{sum} value remains unchanged for a certain predefined generation (G_S).

On the other hand, we specify the region for dissolving cluster in order to perform locally-concentrated search regarding the determination of dissolving cluster. Specifically, we define Θ angle region around on depot for the dissolution region and dissolve clusters within this angle region. Also, we define the search region having twice the

area of dissolution region and restrict the range of search in order to increase the search effectiveness.

The details of this dissolution is described as below.

Step 1: Judgment for restoring simplified problem to original one

If current generation g satisfied $g \geq G_F$, dissolve every cluster in order to return to original problem and finish this flow.

Step 2: Timing judgement for partially dissolution

If the above two conditions about dissolution timing were satisfied, go to Step 3. Otherwise, finish this flow.

Step 3: Determination of dissolution area for dissolving cluster

Calculate the dissolution area utilizing unit vector $u = (1.0, 0.0)$, depot and $0 \leq \theta \leq 360$.

Step 4: Dissolving cluster within dissolution area

Dissolve every cluster having multiple customers within dissolution area.

3.4 Fixing mechanism

This fixing mechanism has two opposite mode; fixing and unfixing mode and it depends on the condition which mode is performed if this mechanism is called. In particular, we use the ratio of the total number of customers as the switching condition. In the case this ratio is under 0.8, fixing mode gets executed, otherwise, unfixing mode is carried out.

In fixing mode, a couple of customer that are the highest potential as customer's sequence is fixed. Since the fixing customers are excluded from search space, we expect that this fixing would promote the efficiency of search and reduce a waste of search.

Of course, this fixing has a risk to fix a wrong couple of customers and this wrong fixing may cause to be trapped in local optima. Due to reduce this risk to the absolute minimum, the following function is used to select a couple of customers as a fixing sequence.

$$F_{\text{bond}} = \frac{d_{(C_i^k, C_{i+1}^k)}}{d_{(0, M_{(C_i^k, C_{i+1}^k)})} \times 2} \tag{5}$$

where $d_{(a,b)}$ represents the distance between $a-b$, C_i^k indicates the customer that is ith visited in route k, 0 is depot and $M_{(a,b)}$ means the middle point between $a-b$.

(5) consists of two parts; the distance from the middle point and the distance from the depot. Since a lower value of (5) means more appropriate for fixing a couple of customers, the couple having lowest value of (5) is selected.

On the other hand, unfixing mode tries to release the fixing customers' sequence in the limited area. The basic idea of unfixing is the same as cluster dissolution in section3.3. Therefore, this mode unfixes every fixing customers within the area calculated the same as cluster dissolution technique. We expects this unfixing has the effect of reducing the risk caused by a wrong fixing customers' sequence.

Table 1. Instances.

Instance	tai385	triple-tai385-d3	sixth-tai385-d1	sixth-tai385-d6
# customers	385	1155	2310	2310
# depots	1	3	1	6

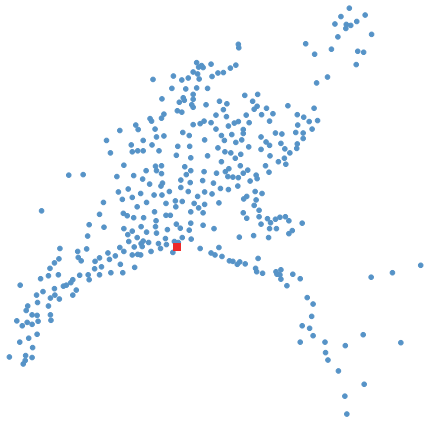


Figure 2. The distribution of tai385.

4 Numerical Examples

We used MOEA/D(A Multiobjective Evolutionary Algorithm Based on Decomposition)(Zhang and Li, 2007) as EMO algorithm. And we treated this MVRPs as two objective problem having F_{dist} and F_{var} of expression (1) and (2) respectively. Table 2 shows parameters we used. As shown in Table 2, the results of this section were of 10 trials.

In this experiment, we investigated the characteristics and effectiveness of the proposed approach by comparing the performance of the approach without our original mechanisms(normal method). In particular, we used four cases for comparison; normal (without every mechanisms), simplification (with simplification and without fixing), immobilization (without simplification and with fixing) and proposed (with every mechanisms) methods.

4.1 VRPs Instances

We used four test problems based on Taillard's instances (tai385) from VRP website;tai385, triple-tai385-d3, sixth-tai385-d1 and sixth-tai385-d6. tai385 is original Taillard's instances and triple-tai385-d3 is created by the combination of 3 tai385. Although sixth-tai385-d1 and sixth-tai385-d6 are composed of 6 tai385, the former instance has only one depot and the latter has 6 depots, respectively.

Table 2. Used Parameters.

Instance	tai385	triple-tai385-d3	sixth-tai385-d1	sixth-tai385-d6
The number of population(A)	50			
The upper period of the search stagnatioG^R	3			
The upper period of the search stagnatio G^B	5			
The number of function call	500,000		1,000,000	
The number of trials	10			

Table 3. Computational times (average).

Problem	Normal	Simplification	Immobilization	Proposed
tai385	23 min 32 s	22 min 30 s	11 min 3 s	10 min 11 s
triple-tai385-d3	12 min 25 s	12 min 36 s	7 min 7 s	9 min 12 s
sixth-tai385-d1	6 h 22 min 46 s	5 h 11 min 11 s	5 h 15 min 25 s	5 h 12 min 59 s
sixth-tai385-d6	1 h 12 min 54 s	55 min 50 s	42 min 46 s	42 min 42 s

Table 4. The values of hyper volume.

Problem	Normal	Simplification	Immobilization	Proposed
tai385	8.15×10^7	8.38×10^7	7.95×10^7	8.22×10^7
triple-tai385-d3	6.44×10^7	6.86×10^7	5.44×10^7	5.66×10^7
sixth-tai385-d1	1.96×10^9	2.62×10^9	1.72×10^9	1.86×10^9
sixth-tai385-d6	2.96×10^8	6.73×10^8	2.13×10^8	6.12×10^8

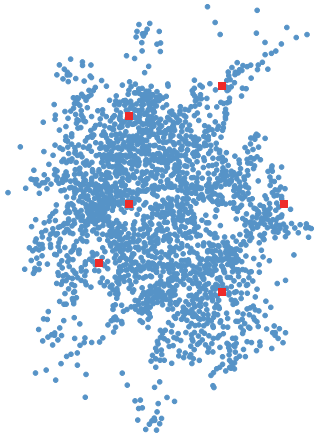


Figure 3. The distribution of sixth-tai385-d6.

The details of four test problems is shown in Table 1 and the distribution of customers in tai385 and sixth-tai385-d6 are shown in Figure 2 and Figure 3.

4.2 Performance measures

In this experience, we used two different type of performance measures for evaluating the obtained solutions; hyper volume (HV) (Zitzler and Thiele, 1998). HV can be calculated as the volume covered by non-dominated solutions and can be treated as an overall measure. In HV, the higher values mean the better solutions.

4.3 Results and Analysis

The results of performance measures are shown in Table 4. The transitions of minimum F_{dist} value in each problem are shown in Figure 4.

From Table 4, we could find that simplification case could get better solutions than those of the other cases in totally. Proposed method obtained better results in sixth-tai385-d6, but in other cases the results of normal method were better than those of proposed method.

However, Figure 4 indicated a bit different result. From these cases, simplification and proposed methods could obtain the solution with better quality. Particularly, in the problem having large customer problems the performance of these methods were overwhelming. Also, in the case of small size problem, normal method were superior to our methods.

From these result, every mechanisms in proposed method have an advantage in large scale problem and are not efficient in the small size problem. Also, simplification could indicate better results in every large scale problem but the performance of immobilization were low in totally. This fact indicates that the only fixing mechanism is difficult to improve a search performance.

On the other hand, Table 3 indicated that every our methods could finish at shorter times than that of normal method. This means every mechanisms of proposed method has an effect on accelerating the search process. Especially, this tendency were significant in large scale problems.

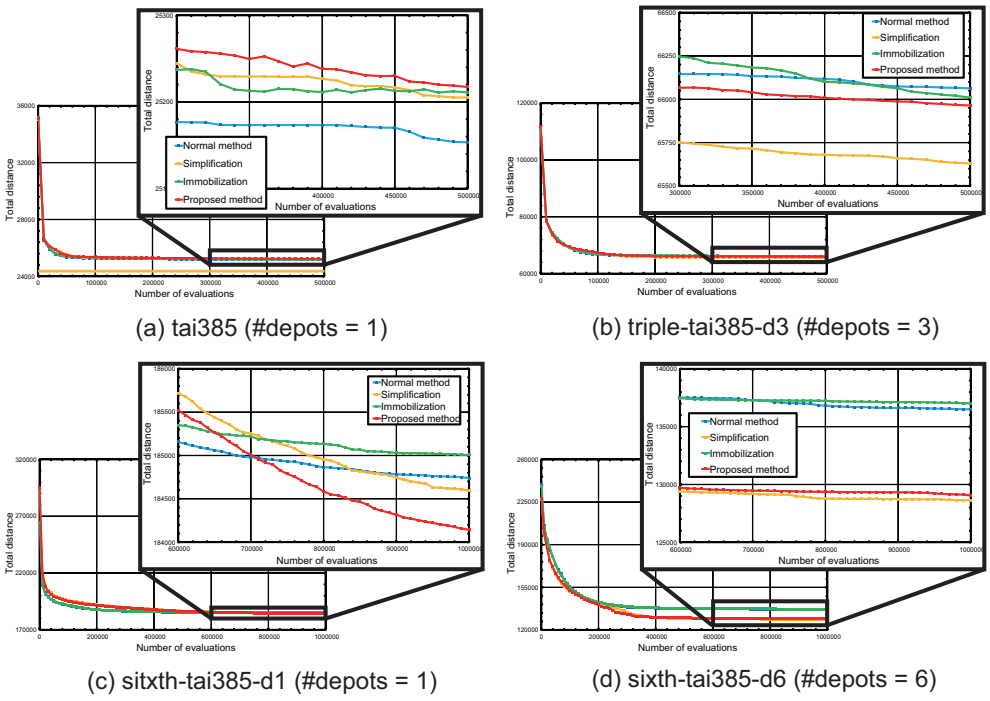


Figure 4. The transitions of minimum F_{dist} value.

5 Conclusions

In this paper, we proposed a new method based on simplification and fixing mechanisms for large scale vehicle routing problems (VRPs). The first mechanism is aimed to obtain high-quality solutions at early stage by mixing simplification and gradual restoration techniques. And the second mechanism is to refine the obtained solution through applying the first mechanism more efficiently due to a reduction of search space.

We investigated the effectiveness of the proposed method and our mechanisms in the proposed method by comparison of its performance with the cases having no our mechanisms.

Numerical experiments clarified the following points:

1) Our mechanisms could obtain the solution with better quality in large customer problems, but in the case of small size problem the proposed method and our mechanisms were not good.
2) To use only fixing mechanism is not effective way for improving a search performance.

As future works, we would investigate the influence of clustering in more detail and try to apply the proposed approach to another very large scale VRPs. Also, we would advance the application to more practical problems (including real world problems).

Acknowledgment

This research was funded in part by a Grant-in-Aid for JSPS fellows (No. 26330269). Also, this work is partially supported by "Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures (jh160047)" and "Information Initiative Center, Hokkaido University (A2-1)" in Japan.

References

O. Braysy and M. Gendreau. Vehicle routing problem with time windows, part i: Route construction and local search algorithms. *Transportation Science*, 39(1):104–118, 2005.

N. Jozefowiez, F. Semet, and E. Talbi. Parallel and Hybrid Models for Multi-objective Optimization: Application to the Vehicle Routing Problem. In *Parallel Problem Solving from Nature—PPSN VII*, pages 271–280, 2002.

J. Luo, X. Li, and M. Chen. Multi-phase meta-heuristic for multi-depots vehicle routing problem. *Software Engineering and Applications*, pages 82–86, 2013.

S. Watanabe and K. Sakakibara. A multiobjectivization approach for vehicle routing problems. In *Evolutionary Multi-Criterion Optimization. Fourth International Conference (EMO 2007),Lecture Notes in Computer Science.*, volume 4403, pages 660–672, 2007.

Q. Zhang and H. Li. Moea/d: A multiobjective evolutionary algorithm based on decomposition. *IEEE Trans. Evolutionary Computation*, 11(6):712–731, 2007.

E. Zitzler and L. Thiele. Multiobjective Optimization Using Evolutionary Algorithms - A Comparative Case Study. *Parallel Problem Solving from Nature - PPSN-V*, pages 292–301, 1998.

Interpolating Lost Spatio-Temporal Data by Web Sensors

Shun Hattori

Web Intelligence Time-Space (WITS) Laboratory, College of Information and Systems, Graduate School of Engineering, Muroran Institute of Technology, 27–1 Mizumoto-cho, Muroran, Hokkaido 050–8585, Japan,
hattori@csse.muroran-it.ac.jp

Abstract

We experience various phenomena (e.g., rain, snow, and earthquake) in the physical world, while we carry out various actions (e.g., posting, querying, and e-shopping) in the Web world. Many researches have tried to mine the Web for knowledge about various phenomena in the physical world, and also several Web services using Web-mined knowledge have been made available for the public. Meanwhile, the previous papers have introduced various kinds of "Web Sensors" with Temporal Shift, Temporal Propagation, and Geospatial Propagation to sense the Web for knowledge about a targeted physical phenomenon, i.e., to extract its spatiotemporal data sensitively by analyzing big data on the Web (e.g., Web documents, Web query logs, and e-shopping logs), and compared them based on their correlation coefficients with Japan Meteorological Agency's physically-sensed spatiotemporal statistics to ensure the accuracy of Web-sensed spatiotemporal data sufficiently. As an industrial application of Web Sensors to a problem of the loss or error of physically-sensed spatiotemporal data due to some sort of troubles (e.g., temporary faults of JMA's observatories), this paper tries to enable Web Sensors to interpolate lost spatiotemporal data of physical statistics by regression analysis

Keywords: spatiotemporal data mining, big data analysis, web sensors, regression analysis

1 Introduction

We experience or forecast various phenomena (e.g., rainfall, snowfall, earthquake, influenza, and traffic accident) in the physical world, while we carry out various actions (e.g., posting, querying, and e-shopping) in the Web world. Recently, there have been many researches to mine a huge amount of various documents in the exploding Web world, especially User Generated Content such as blogs, microblogs (e.g., Twitter), Word-of-Mouth sites, and Social Networking Services (e.g., Facebook), for knowledge about various phenomena and events in the physical world. For instance, opinion and reputation extraction (Dave et al., 2003; Fujimura et al., 2005) of various products and services in the physical world, experience mining (Tezuka et al., 2006; Inui et al., 2008) of various phenomena and events in the physical world, concept hierarchy (semantics) extraction (Hearst, 1992; Ruiz-Casado et al., 2007; Hattori et al., 2008; Hattori and Tanaka, 2008a; Hattori, 2010, 2012a) such as is-a/has-a relationships, and visual appearance (look and feel) extraction (Hattori, 2010; Tezuka and Tanaka, 2006; Hattori et al., 2007; Hattori and Tanaka, 2009; Hattori, 2012b, 2013a) of physical objects in the physical world. Meanwhile, Web services using Web-mined knowledge have been made available for the public, and more and more ordinary people actually utilize them as important information for choosing better products, services, and actions in the physical world.

However, there are not enough investigations (Ginsberg et al., 2009; Sakaki et al., 2010; Aramaki et al., 2011) on how accurately Web-mined data about a targeted phenomenon or event in the physical world reflect physical-world data. It is not so difficult to mine the Web for some kind of potential knowledge data by using various text mining techniques, and it might not be problematic only to enjoy browsing the Web-mined knowledge data. But while choosing better products, services, and actions in the physical world, it must be socially-problematic to idolatrously/immoderately utilize the Web-mined data in public Web services without ensuring their accuracy sufficiently.

The previous papers (Hattori and Tanaka, 2008b; Hattori, 2011a,b, 2012c, 2013b,c,d, 2014, 2015) have introduced various kinds of "Web Sensors" to sense the Web for knowledge about a targeted phenomenon (e.g., rainfall, snowfall, and earthquake) in the physical world, i.e., to extract its spatiotemporal numerical values by analyzing big data on the Web, i.e., various action-based data (e.g., Web documents, Web query logs, and e-shopping logs) in the Web world, and investigated how correlated Web-sensed spatiotemporal data are with physically-sensed spatiotemporal data (e.g., rainfall, snowfall, and earthquake statistics of JMA (Japan Meteorological Agency, 2016)) as shown in Figure 1.

Document-based Web Sensors with "Temporal Shift" (Hattori, 2011a, 2013d) showed that

1. The optimized temporal shift parameter δ of Web Sensors depends on physical phenomena: Not-Shifted Web Sensor whose temporal shift parameter δ is ± 0 gives the highest correlation coefficient (i.e., the Web runs parallel to the physical world) for rainfall, Shifted-to-Future Web Sensor whose temporal shift parameter δ is negative gives the highest correlation coefficient (i.e., the Web leads the physical

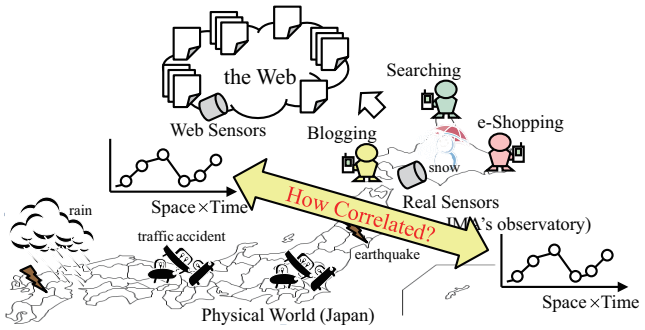


Figure 1. Can Web Sensors sense the physical world sensitively?

world) for snowfall, and Shifted-to-Past Web Sensor whose temporal shift parameter δ is positive gives the highest correlation coefficient (i.e., the Web follows the physical world) for earthquake,

2. The optimized temporal shift parameter δ and correlation coefficient for rainfall are not much dependent on geographical spaces (e.g., 47 prefectures in Japan) and time periods, while the optimized temporal shift parameter δ for snowfall and earthquake varies more widely, and

3. More shaken geographical spaces and time periods are given higher correlation coefficient between Web-sensed spatiotemporal data and physically-sensed spatiotemporal data by the Great East Japan Earthquake (3.11).

Query-based Web Sensors using Web search query logs (Hattori, 2013c) are superior to Document-based Web Sensors using Web documents such as blogs for snowfall and earthquake, while Query-based Web Sensors are inferior to Document-based Web Sensors for rainfall. In addition, the best combined Web Sensor using both Web search query logs and Web documents is superior to uncombined Web Sensors using only Web search query logs or Web documents.

This paper introduces a novel method to interpolate the loss of physically-sensed spatiotemporal data about a targeted physical phenomenon (e.g., Japan Meteorological Agency's rainfall, snowfall, and earthquake statistics) by regression analysis between physically-sensed spatiotemporal data and Web-sensed spatiotemporal data about the targeted physical phenomenon, as an industrial application of variously defined "Web Sensors" with Temporal Shift, Temporal Propagation, and Geospatial Propagation to sense the Web for knowledge about a targeted physical phenomenon, i.e., to extract its spatiotemporal data sensitively by analyzing big data on the Web (e.g., Web documents, Web queries, and e-shopping logs).

The rest of this paper is organized as follows. Section 2 shows various definitions of Web Sensors, and Section 3 introduces a novel method of interpolating lost spatiotemporal data of physical statistics by Web Sensors and regression analysis. And Section 4 concludes this paper.

2 Web Sensors

This section shows various definitions of Web Sensors with Temporal Shift, Temporal Propagation, and Geospatial Propagation to sense the Web for spatiotemporal numerical values dependent on a geographic space (e.g., one of 47 prefectures in Japan) and a time period (e.g., days and weeks in 2011) about a physical phenomenon (e.g., rainfall, snowfall, and earthquake).

First, the simplest and spatiotemporally-normalized Web Sensor (Hattori and Tanaka, 2008b; Hattori, 2013b) by using only Web documents (not Web search query logs (Hattori, 2013c)) with a linguistic name of a geographic space s, e.g., one of 47 prefectures in Japan such as "Hokkaido," a time period t, e.g., one of 365 days or 52 weeks in 2011 such as January 1st (1st day) or from January 1st to 7th (1st week) and from December 24th to 30th (52nd week), and a linguistic keyword kw representing a targeted physical phenomenon, e.g., "rain," "snow," and "earthquake," is defined as

$$\mathrm{ws}(kw,s,t) := \frac{\mathrm{df}_t(\texttt{["}kw\texttt{" AND "}s\texttt{"]})}{\mathrm{df}_t(\texttt{["}s\texttt{"]})}, \qquad (1)$$

where $\mathrm{df}_t(\texttt{["}s\texttt{"]})$ stands for the Frequency of Web Documents searched from the Web, especially the Weblog, by submitting the search query q with the custom time range t to Google Web Search. Note that the Weblog is superior to the whole Web, Twitter, Facebook, and News as a corpus of Web Sensors (Hattori, 2012c).

Secondly, the temporally-shifted Web Sensor (Hattori, 2011a, 2013d) with a "Temporal Shift" parameter δ [day], a geographic space s, a time period t, and a linguistic keyword kw representing a targeted physical phenomenon is defined as

$$\text{ws-ts}_\delta(kw,s,t) := \mathrm{ws}(kw,s,t+\delta). \qquad (2)$$

As shown in Figure 2, Shifted-to-Past Web Sensor for a targeted physical phenomenon (e.g., earthquake) when its Temporal Shift parameter δ is positive (e.g., +14) calculates a numerical value dependent on a geographic space s (e.g., "Hokkaido" prefecture in Japan) and a time period t (e.g., one of 52 weeks in 2011) by using Web documents uploaded δ day(s) after the time period t (i.e., infers the past from the future), while Shifted-to-Future Web Sensor when its Temporal Shift parameter δ is negative (e.g., -14) calculates a numerical value dependent on a geographic space s and a time period t by using Web documents uploaded $|\delta|$ day(s) before the time period t (i.e., infers the future from the past).

Thirdly, the temporally-propagated Web Sensor (Hattori, 2011a) with a "Temporal Propagation" parameter σ_t^2, a geographic space s, a time period t, and

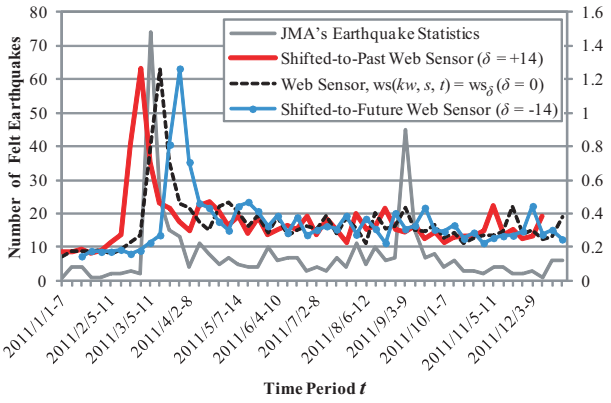


Figure 2. Temporally-shifted Web Sensors for earthquake and JMA's weekly earthquake statistics in Hokkaido prefecture, 2011.

a linguistic keyword kw representing a physical phenomenon is defined by integrating the surrounding time periods as

$$\text{ws-tp}^{\sigma_t^2}(kw,s,t) := \sum_{\forall\delta} \text{ws-ts}_\delta(kw,s,t)\cdot p^{\sigma_t^2}(\delta) \tag{3}$$

$$p^{\sigma_t^2}(\delta) := \frac{1}{\sqrt{2\pi\sigma_t^2}}\cdot\exp\Big(-\frac{\delta^2}{2\sigma_t^2}\Big) \tag{4}$$

where $p^{\sigma_t^2}(\delta)$ stands for a Normal Distribution $N(0,\sigma_t^2,\delta)$ with a mean 0 and a variance σ_t^2. In this paper, $\forall\delta$ is restricted to $[-30,30]$.

Next, the geospatially-propagated Web Sensor (Hattori, 2014, 2015) with a "Spatial Propagation" parameter σ_s^2, a geographic space s, a time period t, and a linguistic keyword kw representing a targeted physical phenomenon is defined by integrating the surrounding geographic spaces as

$$\text{ws-sp}^{\sigma_s^2}(kw,s,t) := \sum_{\forall s_i} \text{ws}(kw,s_i,t)\cdot p^{\sigma_s^2}(\text{distance}(s,s_i)) \tag{5}$$

$$p^{\sigma_s^2}(d) := \frac{1}{\sqrt{2\pi\sigma_s^2}}\cdot\exp\Big(-\frac{d^2}{2\sigma_s^2}\Big) \tag{6}$$

where $\text{distance}(s,s_i)$ stands for the geographic distance [km] between geographic spaces s and s_i and is calculated based on their latitude and longitude. In this paper, $\forall s_i$ is restricted to 47 prefectures in Japan, and the latitude and longitude of its prefectural capital are used for calculating $\text{distance}(s,s_i)$ by using the Survey Calculation API of GSI (GeoSpatial Information Authority of Japan, 2016). In pairs of 47 prefectures in Japan, the pair of Hokkaido pref. (Sapporo city) and Okinawa pref. (Naha city) has the longest distance, 2243.9 [km], while the pair of Shiga pref. (Otsu city) and Kyoto pref. (Kyoto city) has the shortest distance, 10.5 [km].

3 Data Interpolation

As an industrial application of variously above-defined "Web Sensors" with Temporal Shift, Temporal Propagation, and Geospatial Propagation to the loss or error of physically-sensed spatiotemporal data due to some sort of troubles (e.g., temporary faults of Japan Meteorological Agency's observatories), this section proposes a novel method to interpolate lost spatiotemporal data about a targeted physical phenomenon (e.g., Japan Meteorological Agency's rainfall, snowfall, and earthquake statistics).

For a lost spatiotemporal numerical value $\text{ps}(s,t,kw)$ about a targeted physical phenomenon (which is represented by a linguistic keyword kw, e.g., "rain," "snow," and "earthquake") in a geographic space s, e.g., one of 47 prefectures in Japan such as "Hokkaido" over a time period t, e.g., one of 365 days or 52 weeks in 2011 such as January 1st (1st day) or from January 1st to 7th (1st week) and from December 24th to 30th (52nd week), the proposed method interpolates it by regression analysis with its surrounding N physically-sensed spatiotemporal data, their corresponding N Web-sensed spatiotemporal data, and its corresponding Web-sensed spatiotemporal data $\text{ws}(s,t,kw)$ or $\text{ws-XX}(s,t,kw)$ (where $\text{XX}\in\{$"ts," "tp," "sp"$\}$). In this paper, N is restricted to [1,30]. The variety of N physically-sensed spatiotemporal data surrounding a lost physically-sensed spatiotemporal numerical value $\text{ps}(s,t,kw)$ has:

1. N physically-sensed spatiotemporal data followed by it (i.e., only N past data),
$$\text{ps}(s,t-N,kw),\cdots,\text{ps}(s,t-1,kw),$$
2. N physically-sensed spatiotemporal data following it (i.e., only N future data),
$$\text{ps}(s,t+1,kw),\cdots,\text{ps}(s,t+N,kw),$$
3. $\lfloor N/2\rfloor$ physically-sensed spatiotemporal data followed by it and $\lceil N/2\rceil$ physically-sensed spatiotemporal data following it (i.e., both $\lfloor N/2\rfloor$ past data and $\lceil N/2\rceil$ future data, future-preferred when N is odd-numbered),
4. $\lceil N/2\rceil$ physically-sensed spatiotemporal data followed by it and $\lfloor N/2\rfloor$ physically-sensed spatiotemporal data following it (i.e., both $\lceil N/2\rceil$ past data and $\lfloor N/2\rfloor$ future data, past-preferred when N is odd-numbered).

The generalization of the above-mentioned examples is m ($\in[0,N]$) physically-sensed spatiotemporal data followed by it and $N-m$ physically-sensed spatiotemporal data following it (i.e., m past data and $N-m$ future data) as shown in Figure 3. Meanwhile, Figure 4 shows a simple method to interpolate a lost physically-sensed datum by average function using only physically-sensed data.

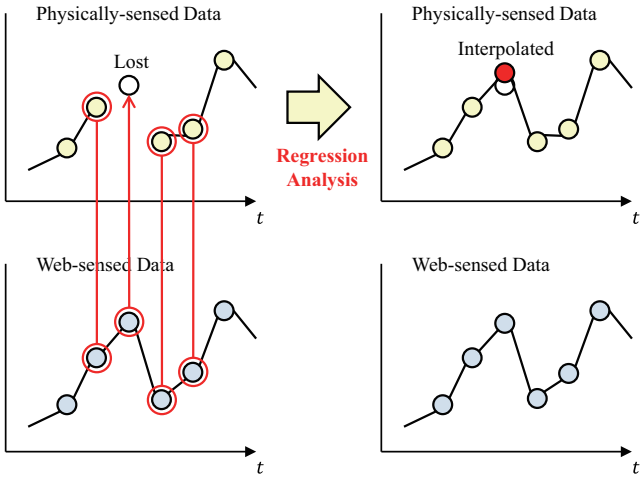


Figure 3. Interpolating a lost physically-sensed datum by Web Sensors and regression analysis using not only physically-sensed data but also Web-sensed data (when $N = 3$ and $m = 1$).

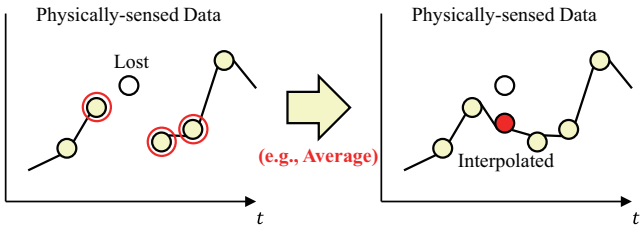


Figure 4. Interpolating a lost physically-sensed datum by average function using only physically-sensed data (adopted as a baseline in the experiment).

4 Conclusions

This paper has introduced a novel method to interpolate the loss of physically-sensed spatiotemporal data about a targeted physical phenomenon (e.g., Japan Meteorological Agency's rainfall, snowfall, and earthquake statistics) by regression analysis between physically-sensed spatiotemporal data and Web-sensed spatiotemporal data about the targeted physical phenomenon, as an industrial application of variously defined "Web Sensors" with Temporal Shift, Temporal Propagation, and Geospatial Propagation to sense the Web for knowledge about a targeted physical phenomenon, i.e., to extract its spatiotemporal data sensitively by analyzing big data on the Web (e.g., Web documents, Web queries, and e-shopping logs).

The future work has to perform experiments to validate the introduced method of interpolating lost spatiotemporal data of physical statistics by Web Sensors and regression analysis, and also will try to apply the other kinds of physical phenomena to the proposed interpolation. In addition, Web Sensors will be able to forecast future data about a targeted physical phenomenon and to alert falsified data of real statistics.

Acknowledgment

This research project was partially supported by a grant-in-aid for scientific research from the Japan Society for Promotion of Science (15K00329).

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pages 1568–1576, 2011.

Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the Peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International World Wide Conference (WWW'03)*, pages 519–528, 2003.

Shigeru Fujimura, Masashi Toyoda, and Masaru Kitsuregawa. A reputation extraction method considering structure of sentence. In *Proceedings of the 16th IEICE Data Engineering Workshop (DEWS'05)*, 6C-i8, 2005.

GeoSpatial Information Authority of Japan. Sokuchi survey calculation API No.2, 2016. http://vldb.gsi.go.jp/sokuchi/surveycalc/main.html.

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, 2009.

Shun Hattori. Object-oriented semantic and sensory knowledge extraction from the Web. In *Web Intelligence and Intelligent Agents*, chapter 18, pages 365–390. In-Tech, 2010.

Shun Hattori. Secure spaces and spatio-temporal Weblog sensors with temporal shift and propagation. In *Proceedings of the 1st IRAST International Conference on Data Engineering and Internet Technology (DEIT'11)*, LNEE vol.157, pages 343–349, 2011a.

Shun Hattori. Linearly-combined Web sensors for spatio-temporal data extraction from the Web. In *Proceedings of the 6th International Workshop on Spatial and Spatiotemporal Data Mining (SSTDM'11)*, pages 897–904, 2011b.

Shun Hattori. Hyponym extraction from the Web based on property inheritance of text and image features. In *Proceedings of the 6th International Conference on Advances in Semantic Processing (SEMAPRO'12)*, pages 109–114, 2012a.

Shun Hattori. Peculiar image retrieval by cross-language Web-extracted appearance descriptions. *International Journal of Computer Information Systems and Industrial Management (IJCISIM)*, 4:486–495, 2012b.

Shun Hattori. Spatio-temporal Web sensors by social network analysis. In *Proceedings of the 3rd International Workshop on Business Applications of Social Network Analysis (BASNA'12)*, pages 1020–1027, 2012c.

Shun Hattori. Hyponymy-based peculiar image retrieval. *International Journal of Computer Information Systems and Industrial Management (IJCISIM)*, 5:79–88, 2013a.

Shun Hattori. Granularity analysis for spatio-temporal Web sensors. In *Proceedings of the WASET International Conference on Knowledge Management (ICKM'13)*, pages 192–200, 2013b.

Shun Hattori. Spatio-temporal Web sensors using Web queries vs. documents. *Journal of Automation and Control Engineering (JOACE)*, 1(3):192–197, 2013c.

Shun Hattori. Spatio-temporal dependency analysis for temporally-shifted Web sensors. In *Proceedings of the 2nd SDIWC International Conference on Informatics & Applications (ICIA'13)*, pages 30–35, 2013d.

Shun Hattori. Spatio-temporal propagation for Web sensors. In *Proceedings of the SDIWC International Conference on Computer Science, Computer Engineering, and Social Media (CSCESM'14)*, pages 69–76, 2014.

Shun Hattori. Deflection analysis for spatially propagated Web sensors. In *Proceedings of the SDIWC International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC'15)*, pages 20–28, 2015.

Shun Hattori and Katsumi Tanaka. Extracting concept hierarchy knowledge from the Web based on property inheritance and aggregation. In *Proceedings of the 7th IEEE/WIC/ACM International Conference on Web Intelligence (WI'08)*, pages 432–437, 2008a.

Shun Hattori and Katsumi Tanaka. Mining the Web for access decision-making in secure spaces. In *Proceedings of the Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on advanced Intelligent Systems (SCIS&ISIS'08)*, TH-G3-4, pages 370–375, 2008b.

Shun Hattori and Katsumi Tanaka. Object-name search by visual appearance and spatio-temporal descriptions. In *Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication (ICUIMC'09)*, pages 63–70, 2009.

Shun Hattori, Taro Tezuka, and Katsumi Tanaka. Mining the Web for appearance description. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA'07)*, LNCS vol.4653, pages 790–800, 2007.

Shun Hattori, Hiroaki Ohshima, Satoshi Oyama, and Katsumi Tanaka. Mining the Web for hyponymy relations based on property inheritance. In *Proceedings of the 10th Asia-Pacific Web Conference (APWeb'08)*, LNCS vol.4976, pages 99–110, 2008.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, volume 2, pages 539–545, 1992.

Kentaro Inui, Shuya Abe, Kazuo Hara, Hiraku Morita, Chitose Sao, Megumi Eguchi, Asuka Sumida, Koji Murakami, and Suguru Matsuyoshi. Experience mining: building a large-scale database of personal experiences and opinions from Web documents. In *Proceedings of the 7th IEEE/WIC/ACM International Conference on Web Intelligence (WI'08)*, pages 314–321, 2008.

Japan Meteorological Agency. Weather, climate & earthquake information, 2016. http://www.jma.go.jp/jma/en/menu.html.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. Automatising the learning of lexical patterns: an application to the enrichment of WordNet by extracting semantic relationships from Wikipedia. *Data & Knowledge Engineering*, 61 (3):484–499, 2007.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International World Wide Web Conference (WWW'10)*, pages 851–860, 2010.

Taro Tezuka and Katsumi Tanaka. Visual description conversion for enhancing search engines and navigational systems. In *Proceedings of the 8th Asia-Pacific Web Conference (APWeb'06)*, LNCS vol.3841, pages 955–960, 2006.

Taro Tezuka, Takeshi Kurashima, and Katsumi Tanaka. Toward tighter integration of Web search with a geographic information system. In *Proceedings of the 15th International World Wide Web Conference (WWW'06)*, pages 277–286, 2006.

Recursive Data Analysis in Large Scale Complex Systems

Esko K. Juuso

Control Engineering, Faculty of Technology, University of Oulu, Finland, esko.juuso@oulu.fi

Abstract

Advanced data analysis is needed in practical applications in large scale complex systems. Variable specific data-driven solutions provide consistent levels, which can be used in compact model structures. In changing operating conditions, the recursive analysis extends the applicability of these structures in building and tuning dynamic and case-based models for complex systems since the meanings change more frequently than the interactions. The methodology provides information about uncertainty, fluctuations and confidence in results. The scaling approach brings temporal analysis to all measurements and features: trend indices are calculated by comparing the averages in the long and short time windows, a weighted sum of the trend index and its derivative detects the trend episodes and severity of the trend is estimated by including also the variable level in the sum. The trend episodes and temporal adaptation of the scaling functions with time are used in the early detection of changes in the operating conditions. The levels are understood as fuzzy labels and the decision making is based on fuzzy calculus. The solution is highly compact: all variables, features and indices are transformed to the range [-2, 2] and represented in natural language which is important in integrating data-driven solutions with domain expertise.

Keywords: recursive data analysis, nonlinear scaling, temporal analysis, fuzzy set systems, large scale systems

1 Introduction

The steady-state simulation models can be relatively detailed nonlinear *multiple input, multiple output (MIMO)* models $\vec{y} = F(\vec{x})$, where the output vector $\vec{y} = (y_1, y_2, \ldots, y_n)$ is calculated by a nonlinear function F from the input vector $\vec{x} = (x_1, x_2, \ldots, x_m)$. More generally, the relationship could also be a table or a graph. Fuzzy set systems, artificial neural networks and neuro-fuzzy methods provide additional methodologies for the function $F(\vec{x})$.

Statistical modelling methodologies provide a wide variety of models based on linear regression. In the *response surface methodology (RSM)*, the relationships are represented with *multiple input, single output (MISO)* models, which contain linear, quadratic and interactive terms (Box and Wilson, 1951). Application areas can be extended by arbitrary nonlinear models, e.g. semi-physical models, developed by using appropriate calculated variables as inputs (Ljung, 1999).

Principal component analysis (PCA) reduces the number of dimensions by using linear combinations of the original variables (Jolliffe, 2002). *Partial least squares regression (PLS)* uses potentially collinear variables (Gerlach et al., 1979). *Nonparametric* models for y_i at each $\vec{x}$ can be constructed from data as a weighted average of the neighbouring values of y_i (Wasserman, 2007)

Fuzzy set systems, which focus on the linguistic meanings of the variables, suit very well to qualitative descriptions of the process as they can be interpreted by using natural language, heuristics and common sense knowledge. Fuzzy logic emerged from approximate reasoning by maintaining clear connections with fuzzy rule-based systems and expert systems (Dubois et al., 1999). *Fuzzy set theory* first presented by Zadeh (Zadeh, 1965) form a conceptual framework for linguistically represented knowledge.

The *extension principle* is the basic generalisation of the arithmetic operations if the *inductive mapping* is a monotonously increasing function of the input. The interval arithmetic presented by Moore (Moore, 1966) is used together with the extension principle for evaluating fuzzy expressions (Buckley and Qu, 1990; Buckley and Hayashi, 1999; Buckley and Feuring, 2000). The fuzzy sets can be modified by intensifying or weakening modifiers (De Cock and Kerre, 2004).

Type-2 fuzzy models take into account uncertainty about the membership function (Mendel, 2007). Most systems based on interval type-2 fuzzy sets are reduced to an interval-valued type-1 fuzzy set. Fuzzy set systems can also handle contradictory data (Krone and Kiendl, 1994; Krone and Schwane, 1996). *Takagi-Sugeno (TS) fuzzy models* (Takagi and Sugeno, 1985) combine fuzzy rules and local lineal models.

Linguistic equation (LE) approach combines data-driven methodologies with linguistic meanings. The LE approach originates from fuzzy set systems (Juuso and Leiviskä, 1992): rule sets are replaced with equations, and meanings of the variables are handled with scaling functions which have close connections to membership functions (Juuso, 1999). The nonlinear scaling technique is needed in constructing nonlinear models with linear equations (Juuso, 2004). Constraints handling (Juuso, 2009) and data-based analysis (Juuso and Lahdelma, 2010), improve possibilities to update the scaling functions recursively (Juuso, 2011a; Juuso and Lahdelma, 2011).

Combined fuzzy systems can include fuzzy arithemetics and inequalities (Juuso, 2014). Natural language in-

terface is based on the scaling functions (Juuso, 2016). Temporal reasoning is a very valuable tool to diagnose and control slow processes: the LE based trend analysis introduced in (Juuso, 2011b) transforms the fuzzy rule-based solution (Cheung and Stephanopoulos, 1990) to an equation-based solution.

Smart adaptive systems (SAS) are aimed for developing successful applications in different fields. Three levels of adaptation have been identified (Anguita, 2001):

1. adaptation to a changing environment;
2. adaptation to a similar setting without explicitly being ported to it;
3. adaptation to a new or unknown application.

In the first level, a short-term memory is needed for incremental or on-line learning, a long-term memory for recognising context drifting. Successful past solutions and the idea of reasoning by analogy are used in the second level. The most challenging requirement is to adapt to new applications. In real applications, the constraint of starting from zero knowledge is modified to building new knowledge or, at least, improving the existing one. Adaptive fuzzy control proceeds through three stages: first scaling, then the shape of membership functions and finally rule-base. The LE approach has a similar preference sequence: scaling, shape of scaling functions and interaction equations.

This paper discusses the recursive data analysis based on the LE approach as a solution in the gradual refinement of large scale complex systems.

2 Data analysis

The nonlinearities of the process are handled by the nonlinear scaling of the variables. The parameters of the scaling functions are obtained by data analysis based on generalised norms and moments.

2.1 Nonlinear scaling

Scaling functions are monotonously increasing functions $x_j = f(X_j)$ where x_j is the variable and X_j the corresponding scaled variable. The function $f()$ consist of two second order polynomials, one for the negative values of X_j and one for the positive values, respectively. The corresponding inverse functions $x_j = f^{-1}(X_j)$ based on square root functions are used for scaling to the range [-2, 2], denoted linguistification. In LE models, the results are scaled to the real values by using the function $f()$.

The parameters of the functions are extracted from measurements by using generalised norms and moments. The support area is defined by the minimum and maximum values of the variable, i.e. the support area is $[\min(x_j), \max(x_j)]$ for each variable $j, j = 1, \ldots, m$. The central tendency value, c_j, divides the support area into two parts, and the core area is defined by the central tendency values of the lower and the upper part, $(c_l)_j$ and $(c_h)_j$, correspondingly. This means that the core area of the variable j defined by $[(c_l)_j, (c_h)_j]$ is within the support area.

The generalised norm is defined by

$$||^{\tau}M_j^p||_p = (M_j^p)^{1/p} = [\frac{1}{N}\sum_{i=1}^{N}(x_j)_i^p]^{1/p}, \qquad (1)$$

where the order of the moment $p \in R$ is non-zero, and N is the number of data values obtained in each sample time τ. The norm (1) calculated for variables x_j, $j = 1, \ldots, n$, have the same dimensions as the corresponding variables. The norm $||^{\tau}M_j^p||_p$ can be used as a central tendency value if all values $x_j > 0$, i.e. $||^{\tau}M_j^p||_p \in R$. (Lahdelma and Juuso, 2011). The norm can be extended to variables including negative values (Juuso, 2011a).

The orders, p, corresponding to the corner points are chosen by using the generalised skewness,

$$(\gamma_k^p)_j = \frac{1}{N\sigma_j^k}\sum_{i=1}^{N}[(x_j)_i - ||^{\tau}M_j^p||_p]^k. \qquad (2)$$

The standard deviation σ_j is the norm (1) with the order $p = 2$. (Juuso and Lahdelma, 2010)

2.2 Interactions

The basic form of the linguistic equation (LE) model is a static mapping in the same way as fuzzy set systems and neural networks, and therefore dynamic models will include several inputs and outputs originating from a single variable (Juuso, 2004). External dynamic models provide the dynamic behaviour, and LE models are developed for a defined sampling interval in the same way as in various identification approaches discussed in (Ljung, 1999).

Adaptation of the nonlinear scaling is the key part in the data-based LE modelling (Figure 1). All variables can be analysed in parallel with the methodology described above and assessed with domain expertise. Interactions are analysed with linear modelling methodologies from the scaled data in the chosen time period. In large-scale systems, a huge number of alternatives need to be compared, e.g. in a paper machine application, 72 variables produced almost 15 million three to five variable combinations. Correlations and causalities based on domain expertise are needed to find feasible variable groups (Juuso and Ahola, 2008).

Dynamic LE models use the parametric model structures, ARX, ARMAX, NARX etc., but the nonlinear scaling reduces the number of input and output signals needed for the modelling of nonlinear systems. For the default LE model, all the degrees of the polynomials become very low:

$$Y(t) + a_1 Y(t-1) = b_1 U(t-n_k) + e(t) \qquad (3)$$

for the scaled variables Y and U .

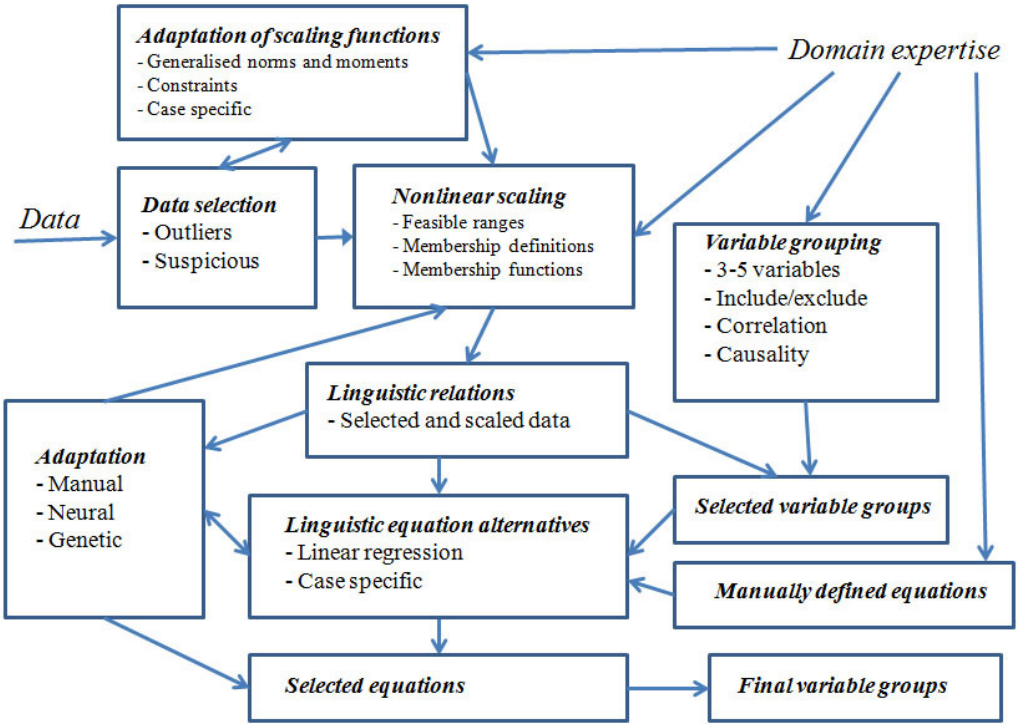


Figure 1. Data-based modelling with linguistic equations (Juuso, 2013).

2.3 Uncertainty

The norm values obtained from different time periods can have differences, i.e. the parameters of the scaling functions can be represented as fuzzy numbers. Thus the feasible range is defined a type-2 trapezoidal membership function. A strong increase in uncertainty may demonstrate a change of operating conditions. The scaling functions monotonous and increasing if the ratios,

$$\begin{aligned} \alpha_j^- &= \frac{(c_l)_j - \min(x_j)}{c_j - (c_l)_j}, \\ \alpha_j^+ &= \frac{\max(x_j) - (c_h)_j}{(c_h)_j - c_j}, \end{aligned} \tag{4}$$

are both in the range $[\frac{1}{3}, 3]$, see (Juuso, 2009).

The coefficients of the second order polynomials can be represented by

$$\begin{aligned} a_j^- &= \tfrac{1}{2}(1-\alpha_j^-)\,\Delta c_j^-, \\ b_j^- &= \tfrac{1}{2}(3-\alpha_j^-)\,\Delta c_j^-, \\ a_j^+ &= \tfrac{1}{2}(\alpha_j^+ - 1)\,\Delta c_j^+, \\ b_j^+ &= \tfrac{1}{2}(3-\alpha_j^+)\,\Delta c_j^+, \end{aligned} \tag{5}$$

where $\Delta c_j^- = c_j - (c_l)_j$ and $\Delta c_j^+ = (c_h)_j - c_j$.

The ratios α_j^- and α_j^- are calculated with interval arithmetic. The constrant range $[\frac{1}{3}, 3]$ must be taken into account before calculating the coefficients (5). Also the extension principle is needed when calculating the scaled values as fuzzy numbers.

2.4 Natural language

All the scaled variables are in the same range [-2, 2] where integer numbers correspond labels, e.g. {very low, low, normal, high, very high}. Fuzzy numbers can be modified by fuzzy modifiers, which are used as intensifying adverbs (very, extremely) or weakening adverbs (more or less, roughly). The resulting terms,

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq A_4 \subseteq A_5, \tag{6}$$

correspond to the powers of the membership in the powering modifiers (Table 1). The vocabulary can also be chosen in a different way, e.g. highly, fairly, quite (Juuso, 2012a). Only the sequence of the labels is important. Linguistic variables can be processed with the conjunction (and), disjunction (or) and negation (not). More examples can be found in (De Cock and Kerre, 2004).

For a time period, the variables are represented by fuzzy numbers whose similarities are compared with the original and modified labels.

Table 1. Modifiers of fuzzy numbers (Juuso, 2016)

Fuzzy number	*Fuzzy label*	*Degree of membership*
A_1	*extremely A*	μ^4
A_2	*very A*	μ^2
A_3	*A*	μ
A_4	*more or less A*	$\mu^{\frac{1}{2}}$
A_5	*roughly A*	$\mu^{\frac{1}{4}}$

3 Recursive analysis

All the phases of the data-based LE modelling shown in Figure 1 can be used in the recursive analysis as well. The recursive part focuses on the scaling and the interactions are updated only if needed.

3.1 Scaling

The parameter of the scaling functions can be recursively updated by using the norms (1) with the defined orders. The norm values are updated by including new equal sized sub-blocks in calculations since the computation of the norms can be done from the norms obtained for the equal sized sub-blocks, i.e. the norm for several samples can be obtained as the norm of the norms of the individual samples:

$$||^{K_s\tau}M_j^p||_p = \{\frac{1}{K_s}\sum_{i=1}^{K_s}[(^{\tau}M_j^p)_i^{1/p}]^p\}^{1/p} \quad (7)$$

$$= [\frac{1}{K_s}\sum_{i=1}^{K_s}[(^{\tau}M_j^p)_i]^{1/p}, \quad (8)$$

where K_s is the number of samples $\{x_j\}_{i=1}^N$. In automation and data collection systems, the sub-blocks are normally used for arithmetic mean ($p = 1$).

Firstly, the parameters of the scaling functions can be recursively updated with by including new samples in calculations. The number of samples can be increasing or fixed with some forgetting or weighting (Juuso, 2011a).

In the second level, the orders of the norms are redefined if the operating conditions change considerably. The new orders are obtained by using the generalised skewness (2) for the data extended with the data collected from the new situation. If the changes are drastic, the calculations are based on the new data only. The decision of starting the redefinition is fuzzy and the data selection is important.

3.2 Interactions

Linear regression and parametric models are used in the recursive tuning of the interaction equations. The set of equation alternatives (Figure 1) is useful in the recursive analysis since the set is validated with domain expertise.

In the first level, the interaction models are not changed. The coefficients are obtained from the data collected from the chosen time period. Uncertainties can be calculated by comparing the coefficients extracted from several short periods.

In the second level, the revised scaling functions may require updates for the interactions as well. However, the re-tuning is started only if the current equations do not operate sufficiently well. The earlier chosen set of alternative equations is used first. New equations are included if new variables become important. The selected variable groups (Figure 1) are analysed first.

Considerable changes in operating conditions mean that the full data-based analysis is needed. This is the third level, which is used to form the model basis for the case-based reasoning (CBR), see (Juuso and Ahola, 2008).

3.3 Fuzzy logic

The recursive data analysis produces parameters for the scaling functions and interactions. Uncertainties of the parameters, which are also obtained in the calculations, are used in detecting changes in operating conditions. The detection is based on fuzzy inequalities $<$, $\leq$, $=$, $\geq$ and $>$ between the new fuzzy parameters and the fuzzy parameters of the case. The resulting a 5X5 matrix includes the degrees of membership of these five inequalities for five parameters. The results are interpreted with the natural language interface which provides an important channel in explaining the changes to the users.

3.4 Smart adaptive systems

The recursive analysis presented above refines the levels of adaptation. The adaptation to a changing environment has two sub-levels: first updating the scaling functions and then interactions if needed. Similar settings are realised with the set of equation alternatives. The adaptation to a new or unknown application includes the full data-based modelling (Figure 1).

4 Temporal analysis

Temporal analysis focused on important variables provides useful information, including trends, fluctuations and anomalies, for decisions on higher level recursive adaptation.

4.1 Trend indices

Trend analysis produces useful indirect measurements for the early detection of changes. For any variable j, a *trend index* $I_j^T(k)$ is calculated from the scaled values X_j with a linguistic equation

$$I_j^T(k) = \frac{1}{n_S+1}\sum_{i=k-n_S}^{k} X_j(i) - \frac{1}{n_L+1}\sum_{i=k-n_L}^{k} X_j(i), \quad (9)$$

which is based on the means obtained for a short and a long time period, defined by delays n_S and n_L, respectively. The index value is in the linguistic range $[-2, 2]$, representing the strength of both decrease and increase of the variable x_j. (Juuso, 2011b; Juuso et al., 2009)

An increase is detected if the trend index exceed a threshold $I_j^T(k) > \varepsilon_1^+$. Correspondingly, $I_j^T(k) < -\varepsilon_1^-$ for a decrease (Figure 2). The derivative of the index $I_j^T(k)$, denoted as $\Delta I_j^T(k)$, is used for analysing the full set of the triangular episodic representations. Trends are linear if the derivative is close to zero: $-\varepsilon_2^- < \Delta I_j^T(k) < -\varepsilon_2^+$. The concave upward monotonic increase (D) and the concave downward monotonic decrease (B) are dangerous situations, which introduce warnings and alarms. The concave downward monotonic increase (A) and the concave upward monotonic decrease (C) mean that an unfavourable trend is stopping.

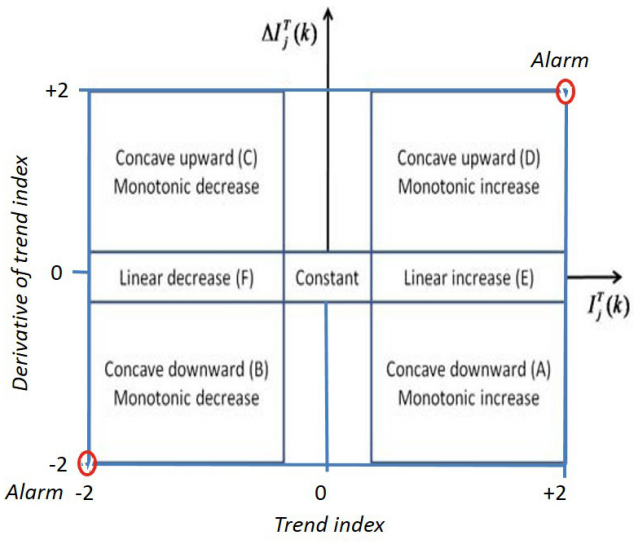


Figure 2. Triangular episodic representations defined by the index $I_j^T(k)$ and the derivative $\Delta I_j^T(k)$.

Severity of the situation can be evaluated by a *deviation index*

$$I_j^D(k) = \frac{1}{3}(X_j(k) + I_j^T(k) + \Delta I_j^T(k)). \qquad (10)$$

This index has its highest absolute values, when the difference to the set point is very large and is getting still larger with a fast increasing speed (Juuso et al., 2009). This can be understood as a third dimension in Figure 2.

The trend analysis is tuned to applications by selecting the time periods n_L and n_S. Further fine-tuning can be done by adjusting the weight factors w_j^{T1} and w_j^{T2} used for the indices $I_j^T(k)$ and $\Delta I_j^T(k)$. The thresholds $\varepsilon_1^+ = \varepsilon_1^- = \varepsilon_2^+ = \varepsilon_2^- = 0.5$. The calculations are done with numerical values and the results are represented in natural language.

The trend analysis can be used for the parameters of the scaling functions and interaction coefficients. Trend of the parameters α_j^-, α_j^-, Δc_j^- and Δc_j^+ give useful information about changes of the scaling functions. The ranges [-2, 2] are $[\frac{1}{3}, 3]$, $[\frac{1}{3}, 3]$. $[c_j - \min(x_j), \max(x_j) - c_j]$, respectively.

4.2 Fluctuations

The *fluctuation indicators*, which were introduced to detecting cloudiness and oscillations, are important improvements aimed for practical use. The indicator calculates the difference of the high and the low values of the corrected irradiation as a difference of two moving generalized norms:

$$\Delta x_j^F(k) = ||^{K_s\tau}M_j^{p_h}||_{p_h} - ||^{K_s\tau}M_j^{p_l}||_{p_l}, \qquad (11)$$

where the orders $p_h \in \Re$ and $p_l \in \Re$ are large positive and negative, respectively. The moments are calculated from the latest $K_s + 1$ values, and an average of several latest values of $\Delta x_j^F(k)$ is used as an indicator of fluctuations. (Juuso, 2012b)

4.3 Changes of operating conditions

Changes of the scaling functions and interaction coefficients are symptoms of changes in operation. The intelligent trend analysis provides early warning about changes in variable levels, fluctuations and uncertainty. All the variables and intelligent indices are represented in the same range [-2, 2], i.e. the same analysis and linguistic interpretation can be applied in all of them. The corresponding levels and their degrees of membership can be used in the fuzzy decision making.

The full analysis is needed fairly seldom although the process changes considerably. For example, new phenomena activate with time in wearing, but the models used in prognostics can be updated by expanding the scaling functions (Figure 3).

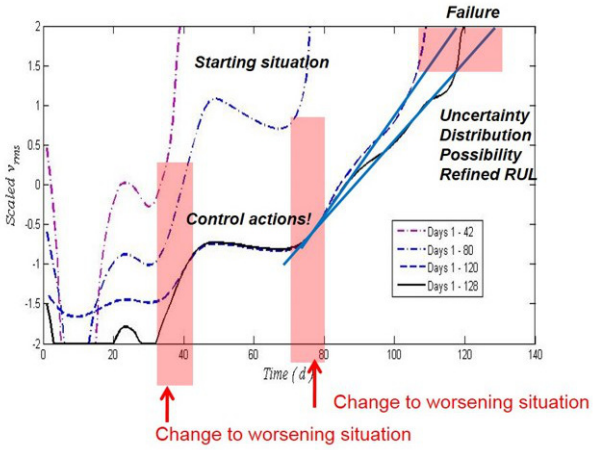


Figure 3. Recursive adaptation in prognostics (Juuso, 2015).

5 Conclusions

The nonlinear scaling approach is the main part of the data processing chain which is the integrating part of the natural language interface. The calculations are done in numeric forms, but the levels and all the indices based on them can be represented in natural language. Data-driven extraction of variable meaning and a set of compact models, with not too many variables in individual equations, form the basis for using the recursive data analysis in practical applications. Included uncertainty representations and the natural language interface help in combining the data analysis with the domain expertise. Complexity needs to be reduced and used in a gradually refining way in practical applications. The recursive data analysis also provides more refined steps for the development of smart adaptive systems.

Acknowledgment

The author would like to thank the research program Measurement, Monitoring and Environmental Efficiency Assesment (MMEA) funded by the TEKES (the Finnish Funding Agency for Technology and Innovation).

References

D. Anguita. Smart adaptive systems - state of the art and future directions for research. In *Proceedings of Eunite 2001 - European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems, July 13-14, 2001, Tenerife, Spain*, pages 1–4. Wissenschaftsverlag Mainz, Aachen, 2001.

G. E. P. Box and K. B. Wilson. On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society. Series B*, 13(1):1–45, 1951.

J. J. Buckley and T. Feuring. Universal approximators for fuzzy functions. *Fuzzy Sets and Systems*, 113:411–415, 2000.

J. J. Buckley and Y. Hayashi. Can neural nets be universal approximators for fuzzy functions? *Fuzzy Sets and Systems*, 101:323–330, 1999.

J. J. Buckley and Y. Qu. On using α-cuts to evaluate fuzzy equations. *Fuzzy Sets and Systems*, 38(3):309–312, 1990.

J. T.-Y. Cheung and G. Stephanopoulos. Representation of process trends - part I. A formal representation framework. *Computers & Chemical Engineering*, 14(4/5):495–510, 1990.

M. De Cock and E. E. Kerre. Fuzzy modifiers based on fuzzy relations. *Information Sciences*, 160(1-4):173–199, 2004.

D. Dubois, H. Prade, and L. Ughetto. Fuzzy logic, control engineering and artificial intelligence. In H. B. Verbruggen, H.-J. Zimmermann, and R. Babuska, editors, *Fuzzy Algorithms for Control, International Series in Intelligent Technologies*, pages 17–57. Kluwer, Boston, 1999.

R. W. Gerlach, B. R. Kowalski, and H. O. A. Wold. Partial least squares modelling with latent variables. *Anal. Chim. Acta*, 112(4):417–421, 1979.

I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, 2 edition, 2002. 487 pp.

E. Juuso. *Integration of intelligent systems in development of smart adaptive systems: linguistic equation approach*. PhD thesis, University of Oulu, 2013. 258 pp., http://urn.fi/urn:isbn:9789526202891.

E. Juuso and S. Lahdelma. Intelligent scaling of features in fault diagnosis. In *7th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies, CM 2010 - MFPT 2010, 22-24 June 2010, Stratford-upon-Avon, UK*, volume 2, pages 1358–1372, 2010. URL `www.scopus.com`.

E. Juuso and S. Lahdelma. Intelligent trend indices and recursive modelling in prognostics. In *8th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies, CM 2011 - MFPT 2011, 20-22 June 2011, Cardiff, UK*, volume 1, pages 440–450. BINDT, 2011. www.scopus.com.

E. Juuso, T. Latvala, and I. Laakso. Intelligent analysers and dynamic simulation in a biological water treatment process. In I. Troch and F. Breitenecker, editors, *6th Vienna Conference on Mathematical Modelling - MATHMOD 2009, February 11-13, 2009, Argesim Report no. 35*, pages 999–1007. Argesim, 2009. ISBN 978-3-901608-35-3.

E. K. Juuso. Fuzzy control in process industry: The linguistic equation approach. In H. B. Verbruggen, H.-J. Zimmermann, and R. Babuška, editors, *Fuzzy Algorithms for Control, International Series in Intelligent Technologies*, volume 14 of *International Series in Intelligent Technologies*, pages 243–300. Kluwer, Boston, 1999. doi:10.1007/978-94-011-4405-6_10.

E. K. Juuso. Integration of intelligent systems in development of smart adaptive systems. *International Journal of Approximate Reasoning*, 35(3):307–337, 2004. doi:10.1016/j.ijar.2003.08.008.

E. K. Juuso. Tuning of large-scale linguistic equation (LE) models with genetic algorithms. In M. Kolehmainen, editor, *Revised selected papers of the International Conference on Adaptive and Natural Computing Algorithms - ICANNGA 2009, Kuopio, Finland, Lecture Notes in Computer Science*, volume LNCS 5495, pages 161–170. Springer-Verlag, Heidelberg, 2009. doi:10.1007/978-3-642-04921-7_17.

E. K. Juuso. Recursive tuning of intelligent controllers of solar collector fields in changing operating conditions. In S. Bittani, A. Cenedese, and S. Zampieri, editors, *Proceedings of the 18th World Congress The International Federation of Automatic Control, Milano (Italy) August 28 - September 2, 2011*, pages 12282–12288. IFAC, 2011a. doi:10.3182/20110828-6-IT-1002.03621.

E. K. Juuso. Intelligent trend indices in detecting changes of operating conditions. In *2011 UKSim 13th International Conference on Modelling and Simulation*, pages 162–167. IEEE Computer Society, 2011b. doi:10.1109/UKSIM.2011.39.

E. K. Juuso. Integration of knowledge-based information in intelligent condition monitoring. In *9th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies, 12-14 June 2012, London, UK*, volume 1, pages 217–228. Curran Associates, NY, USA, 2012a. URL `www.scopus.com`.

E. K. Juuso. Model-based adaptation of intelligent controllers of solar collector fields. In I. Troch and F. Breitenecker, editors, *Proceedings of 7th Vienna Symposium on Mathematical Modelling, February 14-17, 2012, Vienna, Austria, Part 1*, volume 7, pages 979–984. IFAC, 2012b. doi:10.3182/20120215-3-AT-3016.00173.

E. K. Juuso. Intelligent methods in modelling and simulation of complex systems. *Simulation Notes Europe SNE*, 24(1):1–10, 2014. doi:10.11128/sne.24.on.102221.

E. K. Juuso. Recursive data analysis and modelling in prognostics. In *12th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies, CM*

2015 - MFPT 2015, 9-11 June 2015, Oxford, UK, pages 560–567. BINDT, 2015. URL www.proceedings.com. ISBN: 978-1-5108-0712-9.

E. K. Juuso. Informative process monitoring with a natural language interface. In *2016 UKSim-AMSS 18th International Conference on Modelling and Simulation, 6-8 April, 2016, Cambridge, UK*, pages 105–110. IEEE Computer Society, 2016. doi:10.1109/UKSim.2016.37.

E. K. Juuso and T. Ahola. Case-based detection of operating conditions in complex nonlinear systems. In M. J. Chung and P. Misra, editors, *Proceedings of 17th IFAC World Congress, Seoul, Korea, July 6-11, 2008*, volume 17, pages 11142–11147. IFAC, 2008. doi:10.3182/20080706-5-KR-1001.01888.

E. K. Juuso and K. Leiviskä. Adaptive expert systems for metallurgical processes. In S.-L. Jämsä-Jounela and A. J. Niemi, editors, *Expert Systems in Mineral and Metal Processing, IFAC Workshop, Espoo, Finland, August 26-28, 1991, IFAC Workshop Series, 1992, Number 2*, pages 119–124, Oxford, UK, 1992. Pergamon.

A. Krone and H. Kiendl. Automatic generation of positive and negative rules for two-way fuzzy controllers. In H.-J. Zimmermann, editor, *Proceedings of the Second European Congress on Intelligent Technologies and Soft Computing - EUFIT'94, Aachen, September 21 - 23, 1994*, volume 1, pages 438–447, Aachen, 1994. Augustinus Buchhandlung.

A. Krone and U. Schwane. Generating fuzzy rules from contradictory data of different control strategies and control performances. In *Proceedings of the Fuzz-IEEE'96, New Orleans, USA*, pages 492–497, 1996.

S. Lahdelma and E. Juuso. Signal processing and feature extraction by using real order derivatives and generalised norms. Part 1: Methodology. *The International Journal of Condition Monitoring*, 1(2):46–53, 2011. doi:10.1784/204764211798303805.

L. Ljung. *System Identification - Theory for the User*. Prentice Hall, Upper Saddle River, N.J., 2nd edition, 1999.

J. M. Mendel. Advances in type-2 fuzzy sets and systems. *Information Sciences*, 177(1):84–110, 2007.

R. E. Moore. *Interval Analysis*. Prentice Hall, Englewood Cliffs, NJ, 1966.

T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, 15(1):116–132, 1985.

L. Wasserman. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer, Berlin, corr. 3rd edition, 2007.

L. A. Zadeh. Fuzzy sets. *Information and Control*, 8(June): 338–353, 1965.

A Novel Flower Pollination Algorithm based on Genetic Algorithm Operators

Allouani Fouad [1] Kai Zenger [2] Xiao-Zhi Gao [3, 4]

[1]Department of Industrial Engineering, University of Khenchela, Algeria, fouad.allouani@g.enp.edu.dz
[2]Department of Electrical Engineering and Automation, Aalto University, Aalto, Finland, kai.zenger@aalto.fi
[3]Machine Vision and Pattern Recognition Laboratory, Lappeenranta University of Technology, Lappeenranta, Finland.
[4]School of Computing, University of Eastern Finland, Kuopio, Finland, xiao.z.gao@gmail.com

Abstract

The Flower Pollination Algorithm (FPA) is a new natural bio-inspired optimization algorithm that mimics the real-life processes of the flower pollination. Thus, the latter has a quick convergence, but its population diversity and convergence precision can be limited in some applications. In order to improve its intensification (exploitation) and diversification (exploration) abilities, we have introduced a simple modification in its general structure. More precisely, we have added both Crossover and Mutation Genetic Algorithm (GA) operators respectively, just after calculating the new candidate solutions and the greedy selection operation in its basic structure. The proposed method, called FPA-GA has been tested on all the CEC2005 contest test instances. Experimental results show that FPA-GA is very competitive.

Keywords— flower pollination algorithm, crossover, mutation, genetic algorithm (GA)

1 Introduction

Swarm intelligence (SI) optimization algorithms which are inspired by simulation of various types of biological behavior existing in nature, have characteristics of simple operation, good optimization performance and strong robustness. In the last two decades, a large number of algorithms based on this aspect have been suggested, such as, ant colony optimization (ACO) (Socha and Dorigo, 2008), differential evolution (DE) (Storn and Price, 1997), particle swarm optimization(PSO) (Kennedy and Eberhart, 1995), firefly algorithm (FA) (Xinshe, 2012), glowworm swarm optimization (GSO) (Yongquan and Jiakun, 2012), monkey search (MS) (Mucherino and Seref, 2007), harmony search (HS) (Geem *et al*, 2001), cuckoo search (CS) (Yang and Deb, 2009), bat algorithm (BA) (Yang, 2010). SI optimization algorithm can solve complex optimization problems, which classical methods cannot handle efficiently. They have shown excellent performance in many ways (Blum and Li, 2008), and their fields of application are continuously growing (Yang *et al*, 2013).

Flower pollination algorithm (FPA) is a simple and effective SI optimization algorithm proposed in (Yang, 2012). It derives its inspiration from pollination process of flowering plants. From the biological evolution point of view, the objective of flower pollination is the survival of the fittest and the optimal reproduction of plant species. All these factors involved in this process interact systematically between them to achieve optimal reproduction of the flowering plants.

In reality (in the general sense), there are two different ways of pollination; Self-pollination and cross-pollination (Yang, 2012). The cross-pollination (or global pollination) means that pollination can be achieved through pollinators, which carry pollen of a flower of a different plant using Levy flights (Yang, 2012).

The second type, self-pollination (local pollination), is made by the same plant or flower without pollinators. In the latter, the carrying process of pollen is generally done with the help of environmental factors such as wind and diffusion in the water (Yang, 2012).

In this paper, a novel FPA based on both Crossover and Mutation Genetic Algorithm (GAs) operators has been proposed. The changes made allow the introduction of two major improvements: (i) enhancing the diversity of the population, and (ii) improving the intensification ability by the association of these two operators and the elite selection mechanism. Indeed, to demonstrate the efficacy of the proposed algorithm an experimental investigation was carried out using the CEC2005 test suite benchmark problems (Suganthan *et al*, 2005). In addition, the proposed method was also compared to a set of state-of-the-art algorithms including, the basic FPA, the MGOFPA (Draa, 2015), which is a recently proposed FPA variant, the Co-variance Matrix Adaptation Evolution Strategies (CMA-ES) algorithm (Hansen and Ostermeier, 2001), the Comprehensive Learning Particle Swarm Optimizer (CLPSO) (Liang *et al*, 2006), JADE (Zhang and Sanderson, 2009), jDE (Brest *et al*, 2012), CoDE (Wang *et al*, 2011) which are all a DE variants. Moreover, Wilcoxon's rank-sum statistical test was carried out at 5 % significance level to judge whether

the results of the proposed algorithm differ from those of the other algorithms in a statistically significant way (Derrac *et al*, 2011). The rest of this paper is organized as follows: Section 2 presents the fundamental principles of the standard FPA. Section 3 contains a brief description of GAs and its crossover and mutation operators. The proposed algorithm is introduced in Section 4. Experimental results are reported in Section 5. Finally, Section 6 concludes this paper.

2 The Flower Pollination Algorithm

The flower pollination algorithm (FPA) is a new population-based optimization technique inspired by the physiological process of mating in plants. More specifically, this algorithm mimics the reproduction of plants of the same kind or other, through the so-called fertilization or pollination of flowers. To better basically understand the principle of this optimization technique, we start by giving a brief description of its biological underpinnings (Yang, 2012).

2.1 Biological underpinnings of the FPA

Generally, everyone knows that the reproduction of almost plants, in its direct and simple meaning, is a result of a pollination operation. Thus, this very important biological process is typically associated with the transfer of a chemical substance called pollen, and such transfer is often linked with some creatures called pollinators such as insects, birds, bats and other animals.

In fact, some pollinators and certain flowers have co-evolved into very specialized flower-pollinator cooperation. For example, some pollination kind cannot be completed successfully without the intervention of a specific type of pollinators. In reality, there are two main forms in the pollination process; the biotic and abiotic pollination. Thus, about 90% of flowering plants belong in the first class, in which the pollen is transferred by a specific pollinator. Concerning the second class, which does not involve using other organisms and employs wind, water or gravity as pollination mediators, we find only 10% of flowering plants.

Pollinators, or sometimes-called pollen vectors, which may be of various kinds like honeybees for example, represent an essential factor in a biotic pollination form. Thus, some pollinators tend to visit exclusively one species of flower; this pollinator behavior is called flower constancy. The latter increases directly the reproduction of the same flower species by maximizing the transfer of flower pollen to the same plants. This is also advantageous for the pollinators, since they will be sure of the availability of nectar supply with a limited memory and minimum cost of learning.

Depending on the availability of pollinators, two types of pollination are considered; self-pollination and cross-pollination. The first pollination type, called also local pollination, occurs when pollen from one flower pollinates the same flower or other flowers of the same plant (Yang, 2012). Contrariwise, cross-pollination also known as global pollination, happens over long distances when pollen is delivered to a flower from a different plant through a direct or indirect intervention of pollinators following the so-called Lévy flight behavior (Pavlyukevich, 2007).

2.2 The FPA

In (Yang, 2012), Yang emulated the characteristic of the biological flower pollination process in flowering plants to develop the algorithm in question, based on four main rules listed as follows:

Rule1: The global pollination process takes place through biotic and cross-pollination, such that the movement of pollinators has the form of the levy flight (Pavlyukevich, 2007).

Rule2: Local pollination process is considered as abiotic and self-pollination.

Rule3: The flower constancy provided by pollinators is equivalent to a reproduction probability proportional to the similarity of two flowers involved in pollination process.

Rule4: The orientation of the global pollination process, towards local or global pollination is controlled by a switch probability $p \in [0,1]$ with a simple prejudice toward local pollination for reasons relating to the approximation of the algorithm to the real case.

The implementation of these rules is based on a simplistic idea said that: each plant has only one flower, and each flower produces only one pollen gamete (Yang, 2012). Thus, this argument means that it is not necessary to distinguish between a pollen gamete, a flower, a plant or a solution to a problem.

The transition to the mathematical formulation of these rules is carried out according to (Yang, 2012) as follows; first, the global pollination processes (Rule 1), and flower constancy (Rule 3) are represented using the following equation:

$$x_i^{t+1} = x_i^t + \gamma L(\lambda)(g_* - x_i^t) \quad (1)$$

where, x_i^t is the pollen i or the solution vector x_i at iteration t, x_i^{t+1} is the generated solution vector at iteration $t+1$, $g_*$is the current best solution. In addition, γ is a scaling factor used to control the step size. $L(\lambda)$ is the Lévy flights-based step size, it corresponds to the strength of the pollination. In reality, pollinators can fly over a long distance with different distance steps; this can be modeled using a Lévy distribution (Pavlyukevich, 2007) according to the following equation:

$$L \sim \frac{\lambda\Gamma(\lambda)\sin(\frac{\pi\lambda}{2})}{\pi}\frac{1}{s^{1+\lambda}} \quad (s \gg s_0 \gg 0). \qquad (2)$$

In this equation, $\Gamma(\lambda)$ is the standard gamma function, and this distribution is valid for large steps $s > 0$.

Then, the local pollination (Rule 2), and the flower constancy (Rule 3) can be represented as follows:

$$x_i^{t+1} = x_i^t + \epsilon(x_j^t - x_k^t) \qquad (3)$$

where x_j^t and x_k^t are pollen gametes obtained from different flowers of the same plant species. Thus, this random subtraction $(x_j^t - x_k^t)$ is used to imitate the flower constancy in a limited neighbourhood. The parameter ϵ is chosen arbitrarily in $[0,1]$ to approximate this selection to a local random walk (Yang, 2012).

Flower pollination processes can occur randomly at all scales, both local and global case. Hence, to emulate this bi-orientation, a switching parameter p chosen randomly in $[0,1]$ (Rule 4) can be effectively used (Yang, 2012).

The standard FPA is summarised in the following:

Algorithm 1. The flower pollination algorithm

1: Objective Function $f(x)$, $x = (x_1, \ldots, x_d)^T$
2: Initialise a population of n_f flowers in random positions
3: Find the best solution g_* in the initial population
4: Define the switch probability $p \in [0, 1]$
5: Initialise the iteration counter $t = 0$
6: **While** $t < t_{max}$ **do**
7: For $i = 1 : n_f$ (all n_f flowers in the population) do
8: If rand < p then
9: Draw a d-dimensional step vector L which obeys a Lévy distribution
10: Do global pollination via (1)
11: Else
12: Draw ϵ from a uniform distribution in [0,1]
13: Randomly chose x_j^t and x_k^t from the population
14: Do local pollination via (3)
15: End if
16: Evaluate the newly generated solution x_i^{t+1}
17: If the newly generated solution is better, replace x_i^t by x_i^{t+1}
18: Update the current best solution g_*
19: $t = t + 1$
18: End for
19: End while

3 Genetic Algorithm

Genetic algorithm (GA) is a search method that employs random choice to guide a highly exploitative search, by maintaining a balance between exploration of the feasible search domain and exploitation of ‘‘good’’ solutions, see (Holland, 1992). A simple GA is comprised of three main operators: reproduction, crossover, and mutation. Reproduction allocates more copies to solutions with better fitness values and thus imposes the survival-of-the-fittest mechanism on the candidate solutions. Crossover combines partially a set of bits and pieces of two or more parental solutions to produce new, possibly better solutions (i.e. offspring).

Many crossover techniques exist, but the key idea of the most of them is based on the following simple concept: two individuals (parents) are randomly selected and recombined with a probability equal to p_C called crossover probability. Indeed, the combination is achieved if the following condition $rand \leq p_C$ is verified, where $rand$ is random number. Otherwise, the two offspring are simply copies of their parents.

Mutation is the occasional random inversion of bit values that generates non-recursive offspring. More precisely, mutation is often the secondary operator performed with a low probability in GAs. One of the most common mutations method is the bit-flip mutation (Sastry *et al*, 2005). In this kind of mutation, each bit in a binary string is altered (from 0 to 1 or the opposite) with a certain probability p_m known as the mutation probability. In reality, mutation operator performs a random walk near to the individual.

In this paper, we integrate these two GAs operators (crossover and mutation) in the FPA structure to improve its performances. The typical crossover and mutation operation is shown in Figure.1.

Figure 1. Crossover and mutation operation.

4 The Proposed Algorithm

The key idea of the proposed algorithm FPA-GA, is including the concept of crossover and mutation operators as successive steps in the basic FPA. These two steps are included just after calculating the new candidate solutions and the greedy selection operation. Thus, the proposed FPA-GA can be described as shown in the pseudo-code of Algorithm 2 below.

Algorithm 2. FPA based on Crossover and Mutation GAs Operators

1: Objective Function $f(x), x = (x_1, \dots, x_d)^T$
2: Initialise a population of n_f flowers in random positions
3: Find the best solution g_* in the initial population
4: Define the switch probability $p \in [0, 1]$
5: Define the crossover and mutation application probability $Prob\_CrMu \in [0, 1]$
6: $Numb_Obj_Ev = nf$n
7: **While** $Numb\_Obj\_Ev < Max\_Obj\_Ev$ **do**
9: For $i = 1 : n_f$ (all n_f flowers in the population) do
10: If rand < p then
11: Draw a d-dimensional step vector L which obeys a Lévy distribution
12: Do global pollination via (1)
13: Else
14: Draw ϵ from a uniform distribution in [0,1]
15: Randomly chose x_j^t and x_k^t from the population
16: Do local pollination via (3)
17: End if
18: Evaluate the newly generated solution x_i^{t+1}
19: If the newly generated solution is better, replace x_i^t by x_i^{t+1}
20: End for
21: $Numb\_Obj\_Ev = Numb\_Obj\_Ev + n_f$
22: If $Numb\_Obj\_Ev < Max\_Obj\_Ev$ **then**
23: Break ;
24: End if
25: if $rand < Prob\_CrMu$ **then**
26: Apply *crossover operator* on the current population Pop_t to generate a new population Pop_{Cross}
27: Apply *mutation operator* on the Pop_{Cross} population to generate a new population Pop_{Muta}
28: Apply an elite selection of n_f individuals from $\boldsymbol{Pop_t \cup Pop_{Cross} \cup Pop_{Muta}}$
29: $Numb\_Obj\_Ev = Numb\_Obj\_Ev + n_f$
30: end if
31: If $Numb\_Obj\_Ev < Max\_Obj\_Ev$ **then**
32: Break ;
33: End if
34: Update the current best solution g_*
35: **End while**

As shown in Algorithm 2, the main algorithmic structure of the conventional FPA is preserved in the proposed FPA-GA; the supplementary part is shown in gray. Indeed, the intervention of these two operators successively is controlled by the following condition; $rand < Prob\_CrMu$, where $rand$ is a random number $\in [0,1]$ and $Prob\_CrMu$ represents the probability of applying these latter.

Consequently, if this condition is checked two additional populations Pop_{Cross} and Pop_{Muta} will be added respectively. Then, an elite selection takes place to chose the n_f best solutions from the new global generated population $Pop_{Glob} = Pop \cup Pop_{Cross} \cup Pop_{Muta}$.

In addition, we note that Algorithm 2 contains two independent control structures form lines 22–24 and lines 31–33, which their purpose is to avoid execution of extra objective function evaluations by the algorithm. It is also to be noted that the number of objective function evaluations $Numb\_Obj\_Ev$ is always incremented while $Numb\_Obj\_Ev < Max\_Obj\_Ev$ by n_f (see line 21 and 29).

It should be noted that, the strong point of our algorithm resides in the fact that it has a simple algorithmic structure compared with other algorithms, which makes its implementation very easy.

5 Experimental Study

In this section, the FPA-GA algorithm is benchmarked on 25 benchmark functions from a CEC2005 special session (Suganthan *et al*, 2005). The benchmark functions used are minimization functions. They can be divided into four groups: unimodal, multimodal, fixed-dimension multimodal, and composite functions (Suganthan *et al*, 2005). The FPA-GA algorithm was run 20 times on each benchmark function. The number of decision variables is N. For each algorithm (FPA-GA and all other algorithms used in comparative study; FPA, MGOFPA, CMA-ES, CLPSO, JADE, jDE and CoDE and each test function, 20 independent runs were conducted with $n \times 100000$ function evaluations. In our experimental studies, the average and standard deviation ($Mean$ and $Std\ Div$) of the function error value $\left(f(\vec{x}) - f(\vec{x}^*)\right)$ were recorded for measuring the performance of each algorithm, where $\vec{x}$ is the best solution found by the algorithm in a run and $\vec{x}^*$ is the global optimum of the test function. All obtained results are given in Table 1 where the best results are marked in gray spaces. Moreover, $Wilcoxon's$ rank-sum statistical test was carried out at 5% significance level to judge whether the results of FPA-GA algorithm differ from those of the other algorithms in a statistically significant way. In addition, $\ominus$ indicates that FPA-GA performs significantly better than the tested algorithm on the specified function a $\oplus$ indicates that FPA-GA performs not as good as the tested algorithm, and a $\odot$ means that the Wilcoxon rank sum test cannot distinguish between the simulation results of FPA-GA and the tested algorithm. All Wilcoxon rank-sum based comparison of different obtained results are summarized in Table 2.

In all simulation tests, we have adapted FPA-GA, FPA and MGOFPA respectively with the following parameters combination: $p = 0.2, n_f = 50$, $Prob\_CrMu = 0.1$ $Prob\_GOB = 0.1$ (Draa, 2015), $p_C = 0.55$, the mutation rate p_m is given by: $p_m = 1 - \frac{0.1}{N_b} \times i, i = 1, \dots \dots, N_b$ is used in the following condition $p_m < p_{m\,rand}$, where $p_{m\,rand}$ is a random number. Furthermore, in this paper we have used the

following crossover and mutation kinds: arithmetical crossover (Michalewicz, 1992) and non-uniform mutation (Michalewicz, 1992).

Consequently, we can observe clearly from these two tables that FPA-GA performed better than all other algorithm. More precisely (see Table 2), the FPA-GA performed better than FPA, MGOFPA, CMA-ES, CLPSO, JADE, jDE and CoDE in 24, 24, 21, 23, 20, 23 and 21 cases (functions) respectively out of 25 and equal to these latter in 1 ,1,1,1,2,1,2 cases out of 25. Also, FPA-GA performs worse in 3, 1,3,1,2 cases out of 25 than CMA-ES, CLPSO, JADE, jDE and CoDE respectively.

It is clear from this simple presentation, that adding these two operators (crossover and mutation) to the main algorithmic FPA structure allows to improve significantly its performance. Thus, this is due primarily to an improvement of the diversity of the population (three sub-populations Pop, Pop_{Cross} and Pop_{Muta}) which greatly increases the chance to find the best solution, and also to an enhancement of the intensification ability by the association of these two operators and the elite selection mechanism.

6 Conclusions

A new hybrid optimisation method named FPA-GA is introduced in this paper, which considerably improves the performance of the original FPA algorithm by integrating the conventional FPA with two GAs main operators; crossover and mutation. In FPA-GA, the aim of using these latter is to improve the diversification and the intensification characteristics. The experimental studies were carried out on 25 global numerical optimization problems used in the CEC2005 special session on real-parameter optimization. FPA-GA was compared with; the standard FPA, a new FPA variant called MGOFPA, the CMA-ES, the CLPSO, and three DE variants called respectively JADE, jDE and CoDE. The obtained experimental results shown clearly that FPA-GA performances are better than the seven competitors.

The proposed FPA-GA algorithm should be used to solve multi-objective optimization problems in the future to validate its performance. In addition, there exists many NP- hard problems in literature, such as traveling salesman problem, graph-coloring problem, finder of polynomials based on root moments (Huang *et al*, 2004) and knapsack problem. In order to test performance of FPA-GA comprehensively, it should be used to solve these NP-hard problems in the future.

References

C. Blum and X. Li. Swarm intelligence in optimization. In *C. Blum, D. Merkle (Eds.), Swarm Intelligence: Introduction and Applications*, Springer Verlag, Berlin, pages 43-86, 2008.

J. Brest, S. Greiner, B. Boskovic, M. Mernik, and V. Zumer. Selfadapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE Trans. Evolut. Comput.,* 10(6):646–657,2006.

A. Draa. On the performances of the flower pollination algorithm- qualitative and quantitative analyses. *Appl. Soft Comput,* 34: 349–371,2015.

J. Derrac, G. Molina, and F. Herrera. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm Evolut Comput*, 1:3–18, 2011.

J. H. Holland. *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control and artificial intelligence*. *MIT Press*, 1992.

D.S. Huang, H.H.S. Ip, and Z. Chi. A neural root finder of polynomials based on root moments. *Neural Comput*. 16(8):1721–1762, 2004.

N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolut. Comput.,* 9(2):159–195, 2001.

Z.W. Geem, J.H. Kim, and G.V. Loganathan. A new heuristic optimization algorithm harmony search. *Simulation*, 76(2):60–68, 2001.

J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of the IEEE International Conference on Neural Networks*, Perth, Australia, pages 1942–1948, 1995.

J. J. Liang, A. K. Qin, P. N. Suganthan, and S. Baskar. Comprehensive learning particle swarm optimizer for global optimization of multimodal functions. *IEEE Trans. Evolut. Comput.,* 10(3): 281–295, 2006.

Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs*, *Springer-Verlag*, New York, 1992.

A. Mucherino and O. Seref. Monkey search: a novel metaheuristic search for global optimization. In *Proceedings of the American Institute of Physics Conference*, USA, pages 162–173,2007.

I. Pavlyukevich. L´evy flights, non-local search and simulated annealing. *J. Computational Physics*, 226: 1830–1844, 2007.

K. Sastry, D. Goldberg, and G. Kendall. Genetic algorithms. In *E.K. Burke and G. Kendall (eds.). Introductory Tutorials in Optimisation, Decision Support and Search Methodology*. ISBN: 0387234608, Springer. Chapter 4, pages 97-125, 2005.

K. Socha and M. Dorigo. Ant colony optimization for continuous domains. *Eur. J. Op. Res,* 185(3): 1155–1173, 2008.

R. Storn and K. Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim,*11:341–359, 1997.

P. N. Suganthan, N. Hansen, J. J. Liang, K. Deb, Y.-P. Chen, A. Auger, and S. Tiwari. *Problem definitions and evaluation criteria for the CEC 2005 special session on real-parameter optimization*. Nanyang Technol. Univ., Singapore, Tech. Rep. KanGAL #2005005, IIT Kanpur, India, May 2005.

Y. Xinshe. Multiobjective firefly algorithm for continuous optimization. *Eng. Comput*. 29(2): 175–184, 2012.

L. Z. Yongquan and Z. G. Jiakun. Leader glowworm swarm

optimization algorithm for solving nonlinear equations systems. *Electr. Rev*. 88(1b): 101–106, 2012.

X.S. Yang and S. Deb. Cuckoo search via Lévy flights. In *Proceedings of World Congress on Nature & Biologically Inspired Computing (NaBIC 2009, India),* IEEE Publications, USA, pages 210–214,2009.

X.S. Yang. A new metaheuristic bat-inspired algorithm. In *J.R. Gonzalez, D. A. Pelta, C. Cruz (Eds.), Nature Inspired Cooperative Strategies for Optimization.* Springer-Ver-lag, Berlin, Germany, pages 65–74, 2010.

X. S. Yang, Z. Cui, R. Xiao, A. H. Gandomi, and M. Karamanoglu. *Swarm Intelligence and Bio-Inspired Computation*. Elsevier, *Waltham, MA*, 2013.

X.S. Yang. Flower pollination algorithm for global optimization. In *Unconventional Computation and Natural Computation,* Springer, Berlin, pages 240–249, 2012.

Y. Wang, Z. Cai, and Q. Zhang. Differential evolution with composite trial vector generation strategies and control parameters. *IEEE Trans. Evol. Comput.*, 15(1): 55–66, Feb. 2011.

J. Zhang and A. C. Sanderson. JADE: adaptive differential evolution with optional external archive. *IEEE Trans. Evolut. Comput.,* 13(5):945-958, 2009.

Table 1. Experimental Results of FPA, MGOFPA, CMA-ES, CLPSO, JADE, jDE, CoDE and FPA-GA over 20 Independent runs on 25 test functions of $n = 30$ variables with 100000 Max_Obj_Ev.

Function	FPA-GA		FPA		MGOFPA		CMA-ES		CLPSO		JADE		jDE		CoDE	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
1	0.00e+00	0.00e+00	4.03e-29	1.08e-28	1.26e-29	4.59e-29	1.80e-25	4.65e-26	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00
2	3.38e-11	7.54e-11	3.44e-18	8.02e-18	1.55e-02	1.79e-02	6.37e-25	1.78e-25	8.12e+02	2.32e+02	9.37e-29	1.07e-28	8.84e-07	1.12e-06	1.77e-15	2.19e-15
3	3.77e-10	1.65e-10	1.38e+06	1.87e+06	1.27e+06	8.95e+05	5.13e-21	1.30e-21	1.70e+07	2.61e+06	6.57e+03	3.64e+03	2.02e+05	9.92e+04	1.08e+05	5.38e+04
4	4.66e-06	1.15e-06	2.46e-04	4.79e-04	3.73e-01	3.87e-01	6.11e+05	1.68e+06	6.79e+03	1.10e+03	2.56e-14	8.55e-14	3.07e-02	5.68e-02	8.09e-03	2.24e-02
5	1.64e-04	1.27e-04	5.92e+01	2.21e+01	4.80e+01	7.00e+00	3.35e-10	8.62e-11	4.13e+03	4.76e+02	7.89e-06	3.52e-05	5.65e+02	5.22e+02	5.14e+02	4.42e+02
6	1.07e-11	4.12e-12	2.16e+01	4.26e+01	1.75e+01	2.00e+01	3.98e-01	1.22e+00	5.90e+00	1.27e+01	1.00e+01	2.85e+01	2.47e+01	2.69e+01	7.39e-10	1.99e-09
7	8.32e-05	5.38e-06	1.85e-02	1.43e-02	2.68e-02	3.52e-02	1.81e-03	4.39e-03	4.48e-01	8.44e-02	8.17e-03	7.32e-03	1.19e-02	7.76e-03	7.41e-03	8.51e-03
8	4.67e-02	1.46e-04	2.10e+01	9.79e-02	2.10e+01	7.65e-02	2.04e+01	6.89e-01	2.09e+01	5.05e-02	2.09e+01	6.48e-02	2.09e+01	3.31e-02	2.01e+01	1.05e-01
9	1.76e-03	1.42e-04	5.27e+01	2.38e+01	2.55e+01	9.76e+00	4.00e+02	1.15e+02	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00
10	1.13e-03	1.47e-04	1.15e+02	8.63e+01	4.36e+01	3.58e+01	4.41e+01	1.49e+01	1.04e+02	1.77e+01	2.25e+01	2.96e+00	5.33e+01	8.70e+00	3.82e+01	1.14e+01
11	2.00e-02	1.04e-03	3.29e+01	8.41e+00	2.12e+01	7.49e+00	6.72e+00	2.23e+00	2.60e+01	1.72e+00	2.54e+01	2.27e+00	2.76e+01	1.46e+00	1.32e+01	3.84e+00
12	2.55e-02	1.54e-03	4.19e+01	7.61e+00	3.19e+01	8.43e+00	1.36e+04	1.42e+04	1.95e+04	5.56e+03	6.30e+03	7.21e+03	6.05e+03	5.30e+03	2.63e+03	1.91e+03
13	5.83e-04	2.58e-04	7.50e+00	7.29e+00	3.91e+00	2.97e+00	3.22e+00	8.63e-01	2.10e+00	2.21e-01	1.51e+00	6.68e-02	1.68e+00	1.21e-01	1.56e+00	3.17e-01
14	6.88e-02	7.40e-04	1.34e+01	5.76e-01	1.35e+01	2.98e-01	1.47e+01	2.33e-01	1.27e+01	2.28e-01	1.22e+01	2.76e-01	1.29e+01	2.20e-01	1.24e+01	4.91e-01
15	7.51e-04	4.35e-05	2.53e+02	8.60e+01	3.00e+02	1.12e+02	3.67e+02	2.10e+02	6.12e+01	4.10e+01	3.51e+02	1.14e+02	3.89e+02	8.78e+01	4.00e+02	7.94e+01
16	7.87e-04	1.05e-04	2.66e+02	1.15e+02	9.27e+01	8.02e+01	3.65e+02	3.10e+02	1.70e+02	3.47e+01	1.28e+02	1.42e+02	7.81e+01	2.23e+01	6.45e+01	1.64e+01
17	7.58e-04	1.29e-04	1.90e+02	1.29e+02	2.68e+02	8.71e+01	4.97e+02	3.47e+02	2.54e+02	4.25e+01	1.18e+02	1.00e+02	1.46e+02	4.27e+01	6.51e+01	1.27e+01
18	6.82e-04	2.44e-05	8.26e+02	2.01e+00	9.04e+02	1.07e+00	9.03e+02	2.12e-01	9.14e+02	1.34e+00	9.03e+02	2.61e-01	9.04e+02	1.23e+00	9.04e+02	9.64e-01
19	6.86e-04	2.96e-05	8.25e+02	1.77e+00	2.12e+02	3.13e+00	9.03e+02	2.12e-01	9.09e+02	2.20e+01	9.04e+02	1.06e+00	9.04e+02	1.04e+00	9.04e+02	1.04e+00
20	6.96e-04	3.14e-05	8.25e+02	2.28e+00	9.03e+02	7.08e-01	9.03e+02	2.99e-01	9.12e+02	8.11e+00	9.04e+02	7.62e-01	9.04e+02	1.07e+00	9.04e+02	1.34e+00
21	6.23e-04	1.69e-05	7.39e+02	1.80e+02	5.33e+02	1.32e+02	5.00e+02	2.63e-12	5.00e+02	1.25e-12	5.00e+02	5.05e-14	5.00e+02	3.91e-14	5.00e+02	8.65e-14
22	4.71e-04	4.94e-05	5.08e+02	4.57e+00	8.74e+02	2.22e+01	8.19e+02	1.30e+01	9.64e+02	1.05e+01	8.66e+02	2.05e+01	8.74e+02	1.54e+01	8.58e+02	2.23e+01
23	6.31e-04	2.32e-05	7.88e+02	1.83e+02	5.90e+02	1.73e+02	5.36e+02	3.89e+00	5.34e+02	2.04e-04	5.54e+02	9.00e+01	5.34e+02	2.59e-04	5.34e+02	4.51e-04
24	6.17e-04	1.53e-05	2.12e+02	3.13e+00	5.77e+02	3.57e+02	2.00e+02	6.18e-14	2.00e+02	1.46e-12	2.00e+02	2.91e-14	2.00e+02	2.91e-14	2.00e+02	2.91e-14
25	5.05e-04	1.05e-05	2.16e+02	6.91e-01	1.56e+03	1.09e+01	2.10e+02	6.05e+00	2.00E+02	1.96E+00	2.13e+02	7.95e-01	2.11e+02	7.31e-01	2.13e+02	9.12e-01

Table 2. Summarised Wilcoxon rank-sum comparisons between the proposed algorithm as reference and FPA, MGOFPA, CMA-ES, CLPSO, JADE, jDE, CoDE

Functions	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	⊖	⊙	⊕
Algorithms																												
FPA	⊖	⊙	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	**24**	**1**	**0**
MGOFPA	⊙	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	**24**	**1**	**0**
CMA-ES	⊙	⊕	⊕	⊖	⊕	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	**21**	**1**	**3**
CLPSO	⊙	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊕	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	**23**	**1**	**1**
JADE	⊙	⊙	⊖	⊕	⊕	⊖	⊖	⊖	⊕	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	**20**	**2**	**3**
jDE	⊙	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊕	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	**23**	**1**	**1**
CoDE	⊙	⊕	⊖	⊖	⊖	⊙	⊖	⊖	⊕	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	**21**	**2**	**2**

DOI: 10.3384/ecp171421060

A Search Method with User's Preference Direction using Reference Lines

Tomohiro Yoshikawa

Graduate School of Engineering, Nagoya University, Nagoya, Japan, {yoshikawa}@cse.nagoya-u.ac.jp

Abstract

Recently, a lot of studies on Multi-Objective Genetic Algorithm (MOGA), in which Genetic Algorithm is applied to Multi-objective Optimization Problems (MOPs), have been reported actively. MOGA has been also applied to engineering design fields, then it is important not only to obtain Pareto solutions having high performance but also to analyze the obtained Pareto solutions and extract the knowledge in the designing problem. The another has studied the analysis methods of acquired Pareto solutions by MOGA. The aim of these methods is, however, to analyze solutions, and the feedback of the analysis results into the search is little focused. This paper proposes a search method that uses reference lines for a user's preference direction which is defined by the user's attention based on visualization results of a pre-search.

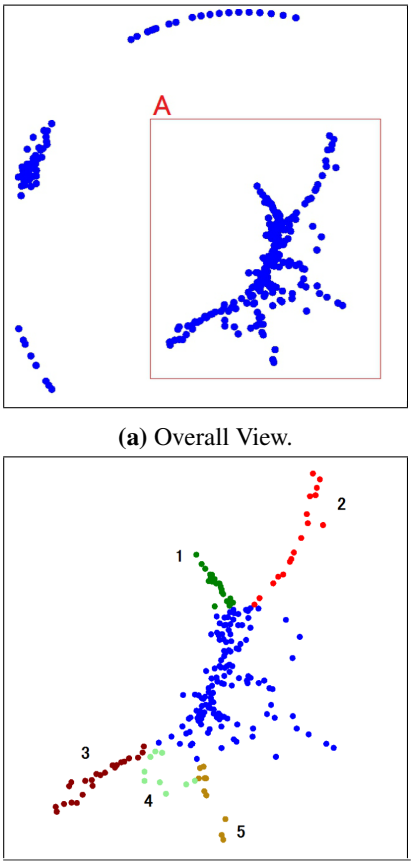
Keywords: optimization problems, multi-objective genetic algorithm, user's preference direction, reference Line, visualization

1 Introduction

In recent years, it is reported that Multi-Objective Genetic Algorithm (MOGA) (Deb, 2001) is applied to engineering design problems in the real-world due to the improvement of computing performance (Obayashi, 2003; Deb, 2003; Oyama and Kawakatsu, 2010). In the engineering design problems, it is required not only to obtain high performance Pareto solutions using MOGA but also to analyze and extract design knowledge in the problem. And in order to analyze Pareto solutions obtained by MOGA, it is required to consider both the objective space and the design variable space.

The author has proposed some analysis methods of acquired solutions by evolutionary computation based on "visualization" (Yamashiro et al., 2006; Ishiguro et al., 2008; Yamamoto et al., 2010; Kudo and Yoshikawa, 2012). Figure 1 and Figure 2 are the examples of visualization for the analysis of Pareto solutions. However, the goal of these approaches is to analyze the acquired solutions. Designers often need better solutions than the acquired ones, more solutions that have the desired features provided by the analysis, or better fitness values on a certain objective function that keeps the other fitness values.

This paper proposes a search method for a user's preference direction based on "reference lines" which is one of the mechanisms in NSGA-III proposed in (Deb and Jain, 2014). In the proposed method, a user selects the preference area in the visualized space by plotting the acquired solutions, and reference points are generated in the selected area. Reference lines are defined by making connections between the reference points and the original point. Moreover, in the proposed method, a user can move the original point based on his/her desired feature of solutions. This paper includes the results off an experiment that applies the proposed method to a real coded multi-objective knapsack problem (Hirano and Yoshikawa, 2013) and examines the ef-

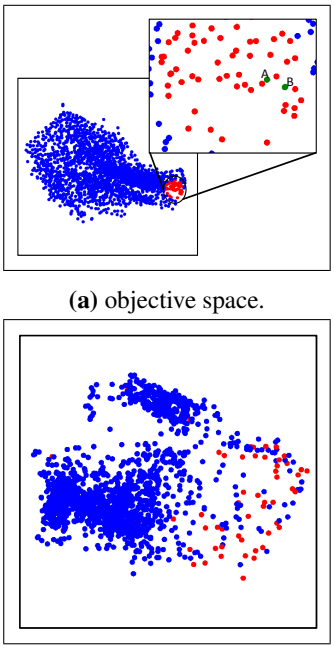


(a) Overall View.

(b) Grouping in each Branch.

Figure 1. Visualization Result using Visualization Tool "ADVICE" in Conceptual Design Optimization Problem of HRE (Hybrid Rocket Engine).

fectiveness of the proposed method.



(a) objective space.

(b) design variable space.

Figure 2. Distribution of Pareto Solutions for Non-Correspondence Area in Trajectory Designing Optimization Problem of "DESTINY".

2 Proposed Method

The proposed method employs the concept of reference lines used in NSGA-III (Deb and Jain, 2014). The aim of the proposed method is to search user's preference direction based on the analysis result generated by the visualization of acquired Pareto solutions. In the proposed method, a user selects the preference area in the visualized space by plotting the acquired solutions, and reference points are generated in the selected area. Reference lines are defined by making connections between the reference points and the original point. Moreover, in the proposed method, a user can move the original point based on his/her desired feature of solutions. As with NSGA-III, the basic algorithm of the proposed method is NSGA-II (Deb, 2002) with the added concept of reference lines. The detail of the operations and the features in the proposed method are described below.

2.1 Selection of Reference Points

In the proposed method, after the analysis generated by the visualization of Pareto solutions acquired by an arbitrary pre-search, a user selects the interested area or individuals based on the analysis results, and then reference points are generated. The search around the selected individuals or highly converged solutions to the user's preference direction can be done by using reference lines connected between the original point shown in section 2.4 and the reference points. There are following two ways to select reference points.

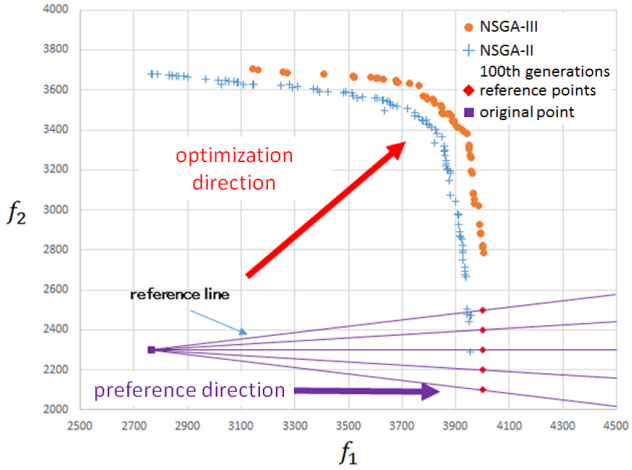


Figure 3. Search by NSGA-III.

- A user selects any individuals directly based on the information of the visualization result of objective space or the fitness values of individuals. The fitness values of these selected solutions are defined as the reference points.

- A user selects interested areas based on the information of the visualization result of objective space. N_r individuals are randomly selected from the selected area and are defined as the reference points.

2.2 Selection of Individuals for Next Generation

NSGA-II and NSGA-III are based on non-dominated sorting. Thus, these methods cannot search well for a user's preference direction when the user's preference direction is far from the center of the optimization direction. Figure 3 shows the acquired Pareto solutions generated by NSGA-III with 100 individuals and 100 generations after 100 generations of NSGA-II, in a two-objective knapsack problem.

In the proposed method, individuals close to the reference lines are selected preferentially for the next generation. It is expected that the number of neighborhood individuals for every reference line becomes same by this section according to the progress of generations. The algorithm for the selection of the next generation is described below.

Step 1: Define the nearest reference line for each individual as a neighborhood line and the individuals belonging to each reference line (neighborhood line) as neighborhood individuals.

Step 2: Select one reference line randomly.

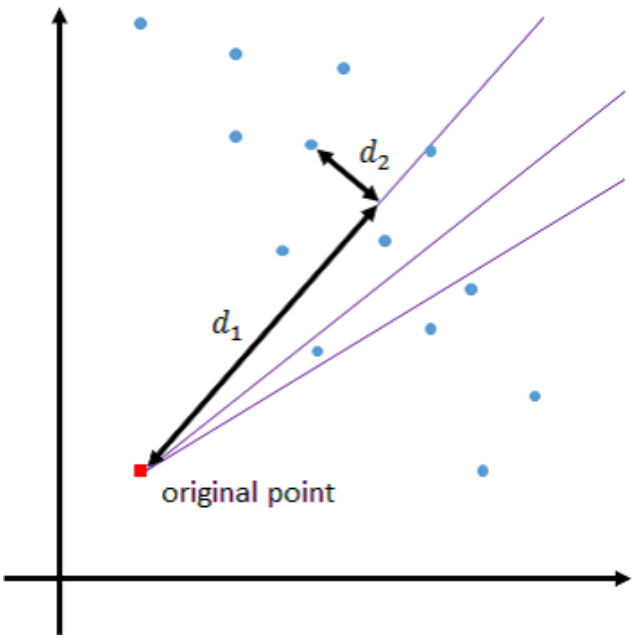


Figure 4. PBI distance.

Step 3: Select the individual that has the largest penalty-based boundary intersection (PBI) distance (Zhang and Li, 2007; Sato, 2014), which is calculated by 1 (see Figure 4), in all neighborhood individuals of the selected reference line for the next generation and remove it from the neighborhood individuals. The PBI distance is used for scalarized optimization, and θ is a penalty parameter. When θ is large, the individuals near the weight vector are given priority. When there is no neighborhood individual at the selected line, select the nearest individual to the line from all remaining individuals for the next generation and remove it.

Step 4: Select other reference lines randomly from the not-selected lines and return to Step 3. If all reference lines are already selected, clear the selected information and return to Step 2.

Step 5: Repeat Step 2 to Step 4 until the number of next generation individuals becomes the population size.

$$\text{PBI distance} = d_1 - \theta d_2 \tag{1}$$

2.3 Selection of Parents for Crossover

In the selection of parent individuals for crossover, tournament selection chooses one parent individual from neighborhood individuals for a reference line. The selected parent individual crossovers with another selected parent individual. The algorithm for the parent selection for crossover is described below.

Step 1: As described in section 2.2, define neighborhood lines and neighborhood individuals.

Step 2: Select one reference line randomly.

Step 3: Depending on the number of neighborhood individuals of the selected reference line, carry out the following selection.

- If the number of neighborhood individuals is more than 1.

  The tournament selection based on PBI distance is performed, *i.e.*, two individuals are randomly chosen and the one that has the larger PBI distance is selected as the parent individual.
- If the number of neighborhood individuals is 1.

  The individual becomes the parent individual.
- If the number of neighborhood individuals is 0.

  The individual closest to the reference line in all individuals becomes the parent individual.

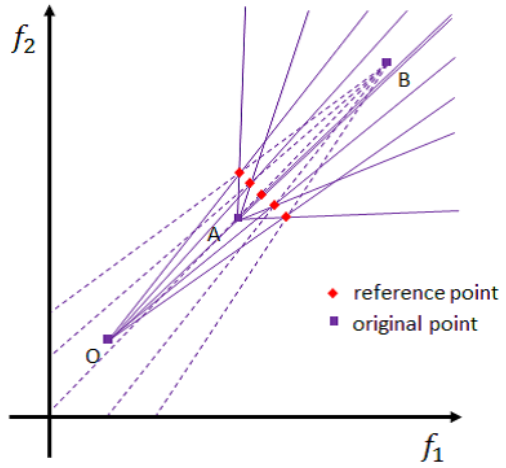
Step 4: Select another parent individual by carrying out Step 2 and Step 3. If the same individual is selected, return to Step 2 until a different one is selected.

2.4 Moving Original Point

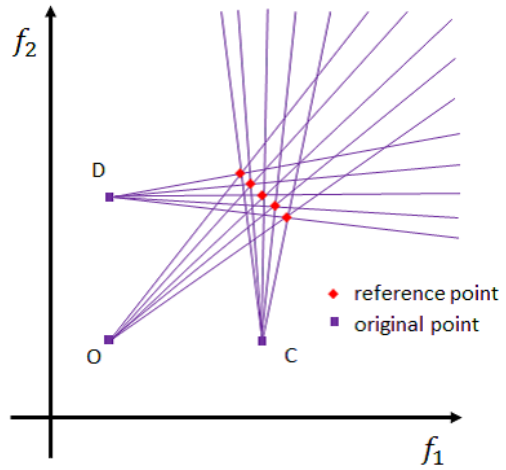
Basically, the original point is set to the worst value of each fitness value in all solutions. However, the proposed method allows a user to move the original point based on his/her desired feature of solutions and change the feature of solutions obtained by the additional search.

Figure 5 shows examples of the original point moving. Figure 5 is an example of maximizing objective functions f_1 and f_2. The original point O is the standard original point described above. Point A in Figure 5(a) is the worst fitness value in the reference points. For example, moving the original point to point A widens the reference lines and it is possible to acquire solutions with large diversity in the user's preference area. Conversely, moving the original point far away from the reference points or to the opposite side of the reference points, such as point B in Figure 5(a) makes the reference lines narrow and the convergence of search solutions is expected to be high or they converge to one point (point B).

Moreover, most fitness values are often satisfied for a user, but the fitness values of specific objective functions are not satisfied in real-world problems. In such cases, by moving the original point such as point C or D in Figure 5(b), the search direction can be concentrated to the user's preference direction. For example, when the original point is set at point D, f_1 will be maximized by keeping f_2.



(a) Adjustment of Width of Searching Individuals.



(b) Adjustment of Weight Objective Functions.

Figure 5. Moving of Original Point.

Thus a user can search preference direction, angle and width easily by moving the original point. For example, if a user wants various individuals, he/she should set the original point at the worst fitness value in the reference points. On the other hand, if a user wants converged individuals, he/she should set the original point at the almost infinite distant point. And if a user wants to keep a certain objective functions and converge others, he/she should move the original point to the center of gravity of reference points in objective functions which he/she wants to keep.

3 Experiment

3.1 Experiment Condition

The experiment described in this paper applied the proposed method to a real coded multi-objective knapsack problem (Hirano and Yoshikawa, 2013) and studied the effectiveness of the proposed method. The number of items was 100. The cost and value of each item were given as random integer values between 0 and 99, and the upper limit of cost was half of the total cost for all items. When the sum of the cost exceeded the upper limit, the exceeded value $\times 5$ was subtracted from the fitness values. Though we do not have to normalize the fitness values in this problem, we should normalize each fitness value when their value scales are different from each other.

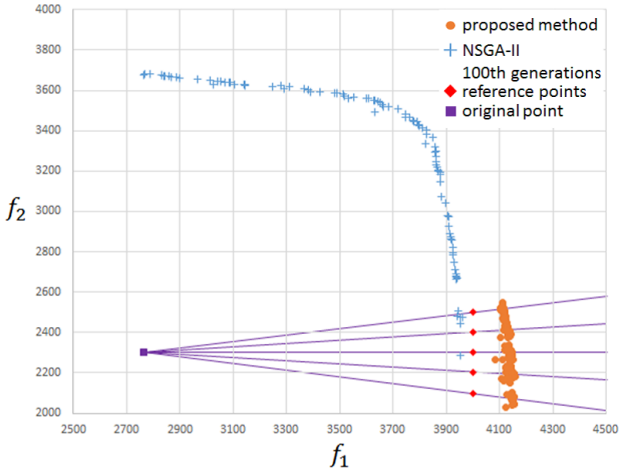


Figure 6. Search by the Proposed Method.

The proposed method assumes that there is a pre-search and that analysis is performed. In this experiment, the pre-search was conducted by NSGA-II. Its population size was 100 and the number of generations was 100. I selected reference points from individuals obtained by pre-search, and searched an additional 100 generations by the proposed method. Then further individuals were gained by pre-search, and added to the initial individuals.

3.2 Effect of Search for Preference Direction

Figure 6 shows the search result by the proposed method using the same conditions as in Figure 3. We can see well-converged solutions for the preference direction which are not present in Figure 3. Although the reference points are usually selected from the acquired solutions in the proposed method, they use the same points as in Figure 3 for the comparison.

3.3 Effect of Moving Original Point

We examined the effect of moving the original point, as described in 2.4. Figure 7 shows the individuals obtained by the proposed method using the original point (a) the worst value of each fitness value $(f1, f2)$ in all individuals including all generations (point O in Figure 5), (b) the worst value of each fitness value in the reference points (point A in Figure 5), (c) infinite distance from the reference points $((f1, f2) = (-100,000, -100,000))$. To evaluate the effectiveness of the proposed method, I use the metric as the performance index. Because the conventional performance indexes are not appropriate to evaluate the performance that this paper desires. Table 1 shows the average distance from the reference point of the neighbor-

Table 1. Convergence and Diversity of Pareto Solutions (2 objectives)

	distance from RP	S.D.	width	S.D.
(a)	136.8	12.8	303.8	46.8
(b)	106.2	11.1	447.8	25.0
(c)	140.8	14.2	276.9	25.0

Table 2. Difference between Center of Gravity of Pareto Solutions and that of Reference Points (2 objectives)

f_1	f_2
168.2	16.7

hood line, which corresponds to the convergence, and the width obtained Pareto solutions, which corresponds to the diversity. Each value in the table is the average of 30 trials and the standard deviation.

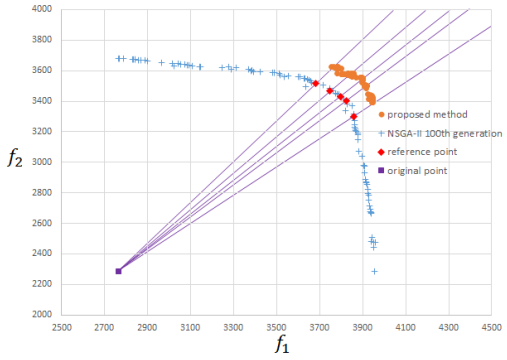
Figure 7 and Table 1 show that the proposed method searched widely by moving the original point close to the reference points, as in Figure 7(b). On the other hand, the proposed method searched narrowly by moving the original point far from the reference points like Figure 7(c), and that made the convergence a little high.

Figure 8 shows the obtained individuals when the position of the original point was set to emphasize f_1 (point D in Figure 7). Moreover, Table 2 shows the difference between the center of gravity of Pareto solutions and that of reference points for each fitness value. Just as intended, the value of f_1 was much improved while that of f_2 remained unchanged.

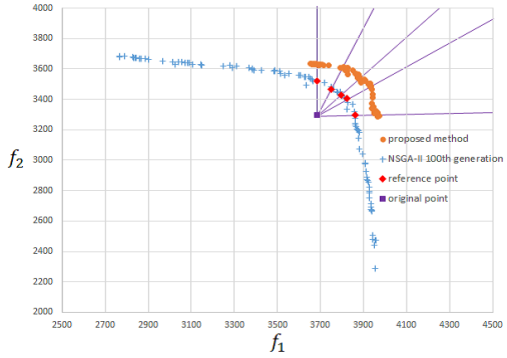
3.4 Effect in Many-objective Optimization Programs

Here, I examine the effect of moving the original point shown in 3.3 in the case of MaOPs. Table 3 shows the same performance indexes (average of 30 trials) with Table 1 in the 2, 4, 6, 8, and 10 objective knapsack problem, respectively. In the real world problems, too many number of objective functions is not practical. So I set the number of objective functions up to 10. This result shows the same tendency as that of Table 1 by moving the original point, in which I can adjust the spread of individuals, even in MaOPs. In MaOPs, the convergence of Pareto solutions acquired by NSGA-II is low; however, the acquired solutions by the proposed method were well-converged and the distance from RP became much larger as the number of objectives increased by concentrating the search area and direction. The selected area and RPs were the same in all trials, and the selected area in 8 objectives might be difficult to search. That could be the reason why the values for 8 objectives were comparatively lower those for 6 objectives.

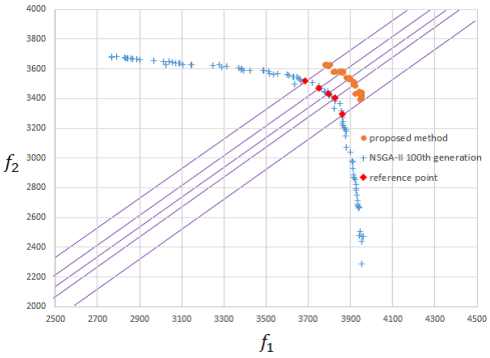
Table 4 shows the result of moving the original point as in Figure 8, in order to search f_1 with priority in $2 \sim 10$ objective problems. The values of Table 4 are the same as those of Table 2. The performance of f_1 is much higher than that of any other objective functions, for all numbers of objectives. This means that moving the original point makes a desired functional search possible in MaOPs as well as 2 objectives.



(a) Worst Fitness Values in All Individuals.



(b) Worst Fitness Values in Reference Points.



(c) Infinite Distance(-100,000, -100,000).

Figure 7. Effect of Moving Original Point.

4 Conclusions

In this paper, I proposed a search method for a user's preference direction by using the concept of reference lines. In the proposed method, a user selects the preference area in the visualized space by plotting the acquired solutions, and reference points are generated in the selected area. Reference lines are defined by making connections between the reference points and the original point. In this

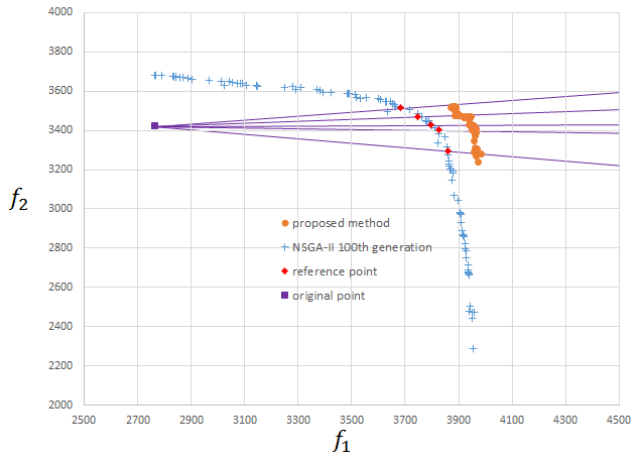


Figure 8. Search with Priority of f_1.

Table 3. Convergence and Diversity of Pareto Solutions (2～10 objectives)

	objectives	2	4	6	8	10
(a)	distance from RP	136.8	575.6	1025.5	682.4	1613.3
	S.D.	12.8	27.9	48.5	40.3	38.9
	width	303.8	1097.9	2614.0	2640.3	3050.9
	S.D	46.8	73.1	179.2	180.3	237.6
(b)	distance from RP	106.2	516.7	865.1	483.7	533.0
	S.D	11.1	27.6	42.2	51.1	51.3
	width	447.8	1871.2	3924.2	4152.1	5045.8
	S.D	25.0	104.2	179.9	241.5	234.7
(c)	distance from RP	140.8	640.2	1369.4	756.1	1950.2
	S.D	14.2	25.9	52.0	54.4	42.7
	width	276.9	979.5	1924.4	2327.1	2469.1
	S.D	25.0	96.2	124.7	162.7	213.8

paper, I described an experiment that applied the proposed method to a real coded multi-objective knapsack problem and studied the effectiveness of the proposed method. The experimental result showed that the solutions having the desired features could be acquired by moving the original point. The experimental result showed that the proposed method also worked well in MaOPs, as well as for 2 objectives. The effect of the proposed method increased with the number of objectives. Future work includes the investigation of the number of reference points and how to define them from the user's preference area. The application and investigation of the proposed method into real-world problems will be also done.

Acknowledgment

This work was supported by the Grant-in-Aid for Scientific Research (C) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, Grant number: 15K00336.

References

K. Deb. *Multi-objective optimization using evolutionary algorithms*. Wiley, 2001. ISBN 047187339X.

Table 4. Difference between Center of Gravity of Pareto Solutions and that of Reference Points (2～10 objectives)

objectives	2	4	6	8	10
f_1	168.2	662.8	531.4	822	297.2
f_2	16.7	-14.4	99	135.3	-57
f_3	-	-11.5	31.7	45.7	15
f_4	-	42.2	51.9	-34.6	70.8
f_5	-	-	-52.9	-47.9	12.7
f_6	-	-	49.4	26.4	-28.6
f_7	-	-	-	41.4	79.3
f_8	-	-	-	22.8	-71.5
f_9	-	-	-	-	70.9
f_{10}	-	-	-	-	-33.2

K. Deb. A Fast and Elitist Multiobjective Genetic Algorithm : NSGA-II. *Evolutionary Multi-Criterion Optimization*, pages 182–197, 2002.

K. Deb. Unveiling innovative design principles by means of multiple conflicting objectives. *Engineering Optimization*, 35(5): 445–470, 2003.

K. Deb and H. Jain. An evolutionary many-objective optimization algorithm using reference-point based non-dominated sorting approach, part I: Solving problems with box constraints. *IEEE Transactions on Evolutionary Computation (TEVC)*, 18(4):577–601, 2014.

Hiroyuki Hirano and Tomohiro Yoshikawa. A study on two-step search based on pso to improve convergence and diversity for many-objective optimization problems. *2013 IEEE Congress on Evolutionary Computation (CEC)*, pages 1854–1859, 2013.

Hidetaka Ishiguro, Tomohiro Yoshikawa, and Takeshi Furuhashi. Visualization of gene-evaluation value in multi-objective problem and feedback for efficient search. *SCIS & ISIS 2008*, pages 1667–1670, 2008.

Fumiya Kudo and Tomohiro Yoshikawa. Knowledge extraction in multi-objective optimization problem based on visualization of Pareto solutions. *2012 IEEE Congress on Evolutionary Computation (CEC)*, pages 860–865, 2012.

S. Obayashi. Multi objective design optimization of aircraft configuration (in Japanese). *The Japanese Society for Artificial Intelligence*, 18(5):495–501, 2003.

K.H. Akira Oyama and Yasuhiro Kawakatsu. Application of multiobjective design exploration to trajectory design of the next-generation solar physics satellite. *Japanese Society for Evolutionary Computation*, 2010.

Hiroyuki Sato. Inverted pbi in moea/d and its impact on the search performance on multi and many-objective optimization. In *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*, pages 645–652. ACM, 2014.

Masafumi Yamamoto, Tomohiro Yoshikawa, and Takeshi Furuhashi. Study on effect of MOGA with interactive island model using visualization. *2010 IEEE Congress on Evolutionary Computation (CEC)*, pages 4196–4201, 2010.

Daisuke Yamashiro, Tomohiro Yoshikawa, and Takeshi Furuhashi. Visualization of search process and improvement of search performance in multi-objective genetic algorithm. *2006 IEEE Congress on Evolutionary Computation (CEC)*, pages 1151–1156, 2006.

Q. Zhang and H. Li. Moea/d: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on Evolutionary Computation (TEVC)*, 11(6):712–731, 2007.

Effects of Chain-Reaction Initial Solution Arrangement in Decomposition-Based MOEAs

Hiroyuki Sato[1] Minami Miyakawa[2] Keiki Takadama[1]

[1] Graduate School of Informatics and Engineering, The University of Electro-Communications, Japan
[2] Faculty of Computer and Information Sciences, Hosei University (JSPS Research Fellow), Japan

Abstract

For solving multi-objective problems, MOEA/D employs a set of weight vectors determining search directions and assigns one solution for each weight vector. Since the conventional MOEA/D assigns a randomly generated initial solution for each weight vector without considering its position in the objective space, mismatched pairs of initial solution and weight are generated, and it causes inefficient search. To enhance MOEA/D based multi-objective optimization, this work proposes a method arranging randomly generated initial solutions to weight vectors based on positions of their solutions in the objective space. The proposed method is combined with the conventional MOEA/D and MOEA/D-CRU, and their search performances are verified on continuous DLTZ4 benchmark problems with 2-5 objectives and different problem difficulty parameters. The experimental results show that the proposed method improves the search performances of MOEA/D and MOEA/D-CRU especially on problems with the difficulty to obtain uniformly distributed solutions in the objective space.

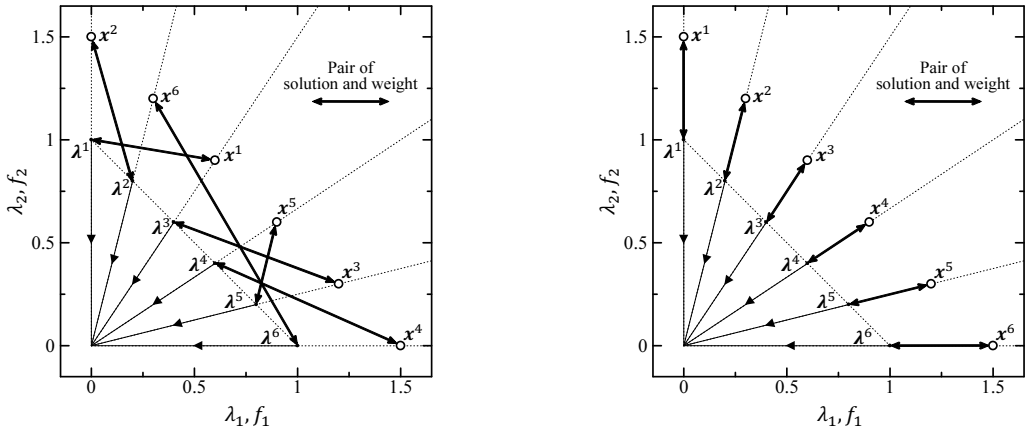
Keywords: multi-objective optimization, many-objective optimization, evolutionary algorithm, MOEA/D

1 Introduction

The aim of multi-objective optimization is to finely approximate the Pareto front, the optimal trade-off among conflicting objectives, with a set of solutions. The population based evolutionary algorithms is a promising approach for multi-objective optimization since a set of solutions to approximate the Pareto front can be picked from the population in a single run (Deb, 2001). Evolutionary multi-objective optimization has been intensively studied so far, and it has been known that the Pareto dominance based NSGA-II (Deb et al., 2002b), SPEA2 (Zitzler et al., 2001), the indicator based IBEA (Zitzler et al., 2004), SMS-EMOA (Beume et al., 2007), HypE (Bader et al., 2011), the decomposition based NSGA-III (Deb et al., 2014), MSOPS (Hughes, 2005), and MOEA/D (Zhang et al., 2007) are representative algorithms. Recently, the decomposition approach is being recognized as an effective approach for solving many-objective problems with more than three objectives. This work focuses on MOEA/D (Zhang et al., 2007) as one of algorithms based on the decomposition approach and tries to improve its search performance.

MOEA/D decomposes a multi-objective problem into a number of single-objective problems and simultaneously optimizes them with a single population. MOEA/D generates a set of weight vectors specifying the decomposition intervals of the objective space. Each weight vector specifies a part of the Pareto front to be approximated. To approximate each part of the Pareto front, MOEA/D assigns one solution for each weight vector. That is, each solution has the role to approximate a part of the Pareto front specified by its weight vector. Before the search, MOEA/D repeats to randomly generate an initial solution for each weight vector. Each initial solution has its own characteristic objective vector and position in the objective space, however, the conventional MOEA/D just sequentially assigns an initial solution to each weight vector without considering its position in the objective space. Consequently, several mismatched pairs of initial solution and weight are generated, and it causes inefficient search. The search performance of MOEA/D would be improved by arranging each initial solution to an appropriate weight vector close to the approximative direction of its initial solution in the objective space.

To improve the search performance of MOEA/D based algorithms, this work proposes a method arranging randomly generated initial solutions to weight vectors based on their positions in the objective space. To arrange the initial solutions, we extend the chain-reaction solution update method (Sato, 2016) previously proposed to update existing solutions in the population with generated offspring. The proposed method assigns each initial solution to a weight vector close to its approximative direction in the objective space. The proposed method makes more appropriate pairs of solution and weight than the conventional method, and the arranged solutions contribute to improving the search performance. This work uses the continuous DTLZ4 benchmark problems with 2-5 objectives and several problem difficulty parameters and verifies the effects of the proposed initial solution arrangement by combined with the conventional MOEA/D (Zhang et al., 2007) and MOEA/D-CRU (Sato, 2016).



(a) Conventional random pairs (b) Proposed arranged pairs

Figure 1. Pairs of solution x and weight λ in the initial population

Algorithm 1 Main MOEA/D Framework (Zhang et al., 2007)

Input: the number of objectives m, the decomposition parameter H, the number of weight vectors and solutions in the population N, the neighborhood size T
Output: the non-dominated set of solutions
1: $\mathscr{L} = \{\lambda^1, \ldots, \lambda^N\} \leftarrow$ Generate weight vectors (H, m)
2: **for each** $\lambda^i \in \mathscr{L}$ **do**
3: $\mathscr{B}_i = \{i_1, \ldots, i_T\} \leftarrow$ Find nearest neighbor weight indices
4: **end for**
5: $\mathscr{P} \leftarrow$ Initialize the population ▷ Algorithm 2 or Algorithm 4
6: **repeat**
7: **for each** $i \in \{1, 2, \ldots, N\}$ **do**
8: $p_1, p_2 \leftarrow$ Randomly select parent indices ($\mathscr{B}_i$)
9: $y \leftarrow$ Generate offspring (x^{p_1}, x^{p_2})
10: Solution Update (y, i) ▷ Algorithm 3 or Algorithm 5
11: **end for**
12: **until** The termination criterion is satisfied
13: **return** The non-dominated solutions picked from $\mathscr{P}$

Algorithm 2 Conventional Population Initialization (Zhang et al., 2007)

1: **procedure** CONVENTIONAL POPULATION INITIALIZATION
2: $\mathscr{P} \leftarrow \emptyset$ ▷ Population
3: **for each** $i \in \{1, 2, \ldots, N\}$ **do**
4: $x^i \leftarrow$ Randomly generate
5: $\mathscr{P} \leftarrow \mathscr{P} \cup x^i$
6: **end for**
7: **return** $\mathscr{P}$
8: **end procedure**

2 MOEA/D

2.1 Algorithm

MOEA/D decomposes a multi-objective optimization problem into a number of single-objective optimization problems and simultaneously optimizes them to approximate the Pareto front of the original multi-objective problem. Algorithm 1 is a pseudocode of the MOEA/D algorithm framework. The original MOEA/D (Zhang et al., 2007) uses Algorithm 2 at 5th line for the population initialization and Algorithm 3 at 10th line for the solution update. In the following, the algorithm of the original MOEA/D is briefly described.

To decompose a m-objective problem, MOEA/D generates $N = C^{m-1}_{H+m-1}$ kinds of weight vectors $\mathscr{L} = \{\lambda^1, \lambda^2, \ldots, \lambda^N\}$ based on the simplex-lattice design with the decomposition parameter H. Each weight vector λ^i specifies a part of the Pareto front to be approximated, its elements $\lambda^i_1, \lambda^i_2, \ldots, \lambda^i_m$ are one of $\{0/H, 1/H, \ldots, H/H\}$, and unique weight vectors satisfying $\sum_{j=1}^{m} \lambda^i_j = 1.0$ are employed for the search. For each weight vector λ^i, a randomly generated solution x^i is assigned, and totally N solutions become the population $\mathscr{P} = \{x^1, x^2, \ldots, x^N\}$. To select parent solutions and update solutions with newly generated offspring, MOEA/D focuses on a weight vector and its neighbor weight vectors. For each weight vector λ^i, its T-neighbors' weight indices $\mathscr{B}_i = \{i_1, i_2, \ldots, i_T\}$ are stored before the search. In the search process, MOEA/D focuses on a weight vector λ^i, selects two parent solutions from solutions assigned to neighbor weights $\mathscr{B}_i$ of the focused weight λ^i, generates an offspring from the selected parents, and tries to update the neighbor solutions with the generated offspring. To compare solutions, a scalarizing function is used. This work employs the weighted Tchebycheff scalarizing function (Li et al., 2014) defined by the following equation.

$$\text{Minimize } g(x|\lambda) = \max_{1 \le j \le m} \{|f_j(x) - z_j| / \lambda_j\}, \qquad (1)$$

where, z is the obtained ideal point, and its each element z_i is the minimum i-th objective value found during the search. $\lambda_j = 0$ is exceptionally replaced with $\lambda_j = 10^{-6}$ to avoid the division by zero.

2.2 Focused Issue

Each weight vector determines a search part of the Pareto front, and its solution assigned to the weight vector tries

Algorithm 3 Conventional Update (Zhang et al., 2007)

```
Input: offspring y, the focused index i
1: procedure CONVENTIONAL UPDATE (y, i)
2:     for each j ∈ B_i do
3:         if g(y|λ^j) is better than g(x^j|λ^j) then
4:             x^j ← y
5:         end if
6:     end for
7: end procedure
```

Algorithm 4 Proposed Population Initialization

```
1: procedure PROPOSED POPULATION INITIALIZATION
2:     Y ← ∅                                  ▷ Temporal population pool
3:     for each i ∈ {1,2,...,N} do
4:         y^i ← Randomly generate
5:         Y ← Y ∪ y^i
6:     end for
7:     P ← ∅                                  ▷ Population
8:     for each i ∈ {1,2,...,N} do
9:         CHAIN-REACTION ARRANGE AND UPDATE (y^i)
10:    end for
11:    return P
12: end procedure
```

to approximate the specified part of the Pareto front. Therefore, for each weight vector, an appropriate solution should be paired are different. However, when the initial solutions are generated, the conventional MOEA/D does not consider pair matchings between solution and weight vector. According to Algorithm 2, the conventional MOEA/D repeats to assign a randomly generated initial solution for each weight vector without considering its position in the objective space. Therefore, several mismatched pairs of solution and weight are generated. Figure 1 (a) shows an example of $N = 6$ initial pairs of solution and weight generated by the conventional MOEA/D with Algorithm 2. This figure shows a two-dimensional weight space ($\lambda_1 - \lambda_2$) and an objective space ($f_1 - f_2$) which both objective functions should be minimized. A set of weight vectors $\lambda^1, \lambda^2, \ldots, \lambda^6$ and a set of initial solutions $x^1, x^2, \ldots, x^6$ are shown in this figure, and six double-headed arrows indicate pair relations of solution and weight vector. In this figure, x^6 has the second worst (highest) value of f_2 among all solutions. However, x^6 is assigned to λ^6 to find the minimum value of f_2 on the Pareto front. If x^4 having the minimum f_2 among all solutions is assigned to λ^6 instead of x^6, the search directed by λ^6 would be enhanced. Since Figure 1 (a) also includes other mismatched pairs of solution and weight, the search to find the Pareto front would be inefficient. Although the initial solutions must be generated randomly, the search would be efficient by arranging each of initial solution to its appropriate weight vector as shown in Figure 1 (b).

3 Proposed Method: Chain-Reaction Initial Solution Arrangement

3.1 Aim and Concept

To improve the search performance of MOEA/D based algorithms by enhancing a number of single objective searches directed by weight vectors, this work proposes a method appropriately arranging randomly generated initial solutions to weight vectors. Since the conventional method just sequentially assigns each initial solution to a weight vector without considering its position in the objective space, and several mismatched pairs of solution and weight are obtained as shown in Figure 1 (a). On the other hand, the proposed method tries to assign each initial solution to an appropriate weight vector with considering its approximative direction, the objective balance vector, in the objective space. Consequently, more appropriate pairs of solution and weight are obtained as shown in Figure 1 (b).

3.2 Method

To arrange the initial solutions, this work extends the chain-reaction solution update method (Sato, 2016). The chain-reaction solution update effectively replaces existing solutions in the population with generated offspring by adaptively determining target existing solutions to be updated based on the position of generated offspring. In the previous work, the chain-reaction solution update is used only for the update of existing solutions with generated offspring. To appropriately arrange initial solutions, this work extends the chain-reaction update, and utilize it in the process at 5th line of Algorithm 1 as an alternative of the conventional Algorithm 2. Algorithm 4 is the pseudocode of the chain-reaction solution arrangement proposed in this work. Algorithm 4 performs Algorithm 5 which is the extended chain-reaction solution update procedure (Sato, 2016) also for the initial solution arrangement. In Algorithm 5, 9-13th lines are newly added to the previously proposed chain-reaction solution update procedure. Since newly added 9-13th lines in Algorithm 5 do not affect to the solution update process, the same Algorithm 5 can be employed also in the solution update process.

The proposed method performs Algorithm 4 at 5th line of Algorithm 1 instead of Algorithm 2. The conventional Algorithm 2 generates the population $\mathscr{P}$ by repeating the random generation of an initial solution to be paired with each weight vector N times. Thus, the conventional method does not care about the positions of randomly generated solutions in the objective space and just sequentially assigns them to weight vectors. On the other hand, to check the relative position and approximative direction of each initial solution in the objective space, the proposed Algorithm 4 first randomly generates N solutions y^i ($i = 1, 2, \ldots, N$) and temporally stores them in the temporal population $\mathscr{Y}$ before assigning them to weight vectors

Algorithm 5 Proposed Chain-Reaction Arrange and Update

Input: solution y
1: **procedure** CHAIN-REACTION ARRANGE AND UPDATE (y)
2: $b \leftarrow$Calculate balance of objective values (y) ▷ Eq. (2)
3: $\mathscr{D} \leftarrow \{d^1, d^2, \ldots, d^N\}$ ▷ Distances from all weight vectors
4: **for each** $i \in \{1,2,\ldots,N\}$ **do**
5: $d^i \leftarrow$ Calculate Euclidean distance (b, λ^i)
6: **end for**
7: $\mathscr{D} \leftarrow$ Sort elements in ascending order $(\mathscr{D})$
8: **for each** $d^j \in \mathscr{D}$ **do**
9: **if** x^j is not exist in the population $\mathscr{P}$ **then**
10: $x^j \leftarrow y$
11: $\mathscr{P} \leftarrow \mathscr{P} \cup x^j$
12: **break**
13: **end if**
14: **if** $g(y|\lambda^j)$ is better than $g(x^j|\lambda^j)$ **then**
15: $tmp \leftarrow x^j$ ▷ Preserve temporally
16: $x^j \leftarrow y$
17: CHAIN-REACTION ARRANGE AND UPDATE (tmp)
18: ▷ Call recursively
19: **break**
20: **end if**
21: **end for**
22: **end procedure**

(3-6 lines). Next, the proposed method makes pairs of solution and weight by repeating Algorithm 5 for each y^i in the temporal population $\mathscr{Y}$ (8-10th lines).

For each initial solution y in the temporal population $\mathscr{Y}$, Algorithm 5 calculates its objective balance vector b by the following equation (2nd line).

$$b_j = \frac{f_j(y) - z_j}{\sum_{\ell=1}^{m} \{f_\ell(y) - z_\ell\}} \quad (j = 1,2,\ldots,m). \tag{2}$$

Next, the proposed method calculates the Euclidean distances between the balance vector b and all weight vectors λ^i $(i = 1,2,\ldots,N)$ and sorts their distances in ascending order (3-7th lines). Then, the proposed method tries to arrange y to a weight vector in short-distance order (8-21th lines). If solution x^j paired with weight vector λ^j which is the closest to the balance vector b still does not exist in the population $\mathscr{P}$, y is assigned to λ^j $(x^j \leftarrow y)$. Otherwise, scalarizing function values of x^j and y are compared in the same way of the chain-reaction solution update (Sato, 2016). If y shows better scalarizing function value than the existing x^j, y becomes new x^j, and Algorithm 5 is recursively performed with the previous x^j like a chain-reaction. For each initial solution y, the above procedure is repeated in the short-distance order of the objective valance vector b and weight vectors until y is assigned to a weight vector.

Since the effects of the proposed solution arrangement depend on the distribution of randomly generated solutions in the objective space, all solutions cannot be assigned to their nearest weight vectors. However, the above procedure can improve the relations between initial solution and weight vector compared with the conventional method. In the case of the example solutions and weight vectors shown in Figure 1, we experimentally verified that pairs of solution and weight becomes the relations shown in Figure 1 (b).

Table 1. MOEA/D based algorithms compared in this work

	Algorithm 1 using	
	Initialization	Solution Update
Conventional MOEA/D	Algorithm 2	Algorithm 3
Proposed MOEA/D-A	Algorithm 4	Algorithm 3
Conventional MOEA/D-CRU	Algorithm 2	Algorithm 5
Proposed MOEA/D-CRU-A	Algorithm 4	Algorithm 5

4 Experimental Settings

To verify the effects of the proposed chain-reaction solution arrangement, this work compares the search performances of four MOEA/D based algorithms. They are the conventional MOEA/D (Zhang et al., 2007) and MOEA/D-CRU (Sato, 2016) without the proposed solution arrangement and the proposed MOEA/D-A and MOEA/D-CRU-A with the proposed solution Arrangement. The differences among four algorithms are described in Table 1.

As the test problems, this work employs DTLZ4 problem framework (Deb et al., 2002a). In the DTLZ test suite, DTLZ4 is the problem framework which can control the difficulty to obtain uniformly distributed non-dominated solutions in the objective space by the problem parameter α. DTLZ4 with $\alpha = 1$ is equivalent to DTLZ2 problem, and $\alpha = 100$ is generally used problem setting for DTLZ4. This work uses 36 patterns of DTLZ4 problems combining $m = \{2,3,4,5\}$ objectives and the parameters $\alpha = \{1,5,10,20,50,100,200,500,1000\}$. Also, the number of variables are set to $n = m + 10$.

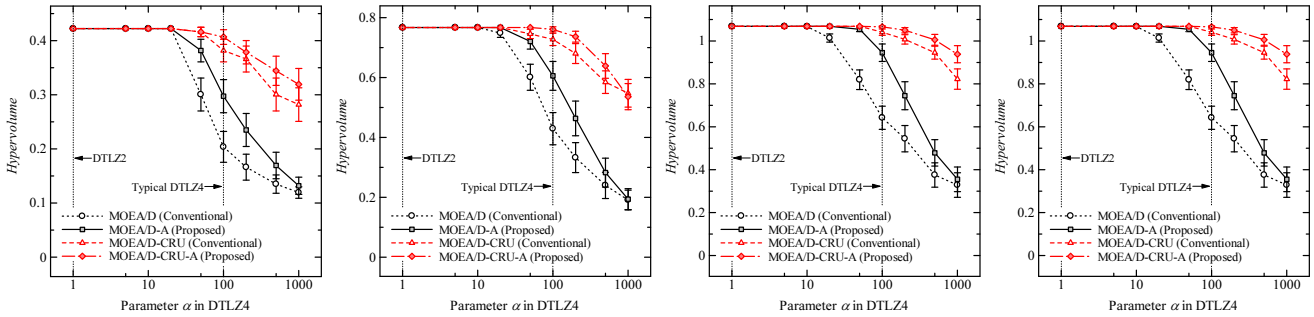
All algorithms employ the decomposition parameters $H = \{200,19,10,7\}$ and the population sizes $N = \{201,210,286,330\}$ for $m = \{2,3,4,5\}$ objective problems, respectively. Also, the neighbor size is set to $T = 20$. To generate offspring solutions, SBX with its ratio 0.8 and distribution index 20 and the polynomial mutation with its ratio $1/n$ and distribution index 20 are used. Also, the termination condition is set to the totally 1,000 generations.

As the search performance metric, this work uses Hypervolume (HV). HV is the m-dimensional volume enclosed by the obtained non-dominated solutions and the reference point $r = (1.1,1.1,\ldots,1.1)$ in the objective space. The higher HV, the higher search performance. In the following experiments, the average HV of 100 independent runs of each algorithm and its 95% confidence intervals are compared.

5 Results and Discussion

5.1 Search Performance at Final Generation

Figure 2 shows the average HV values obtained by the four MOEA/D based algorithms at the final generation. Figure 2 (a)-(d) are results on problems with $m =$



(a) 2 objectives (b) 3 objectives (c) 4 objectives (d) 5 objectives

Figure 2. *HV* at the final generation as the difficulty parameter α is varied

$\{2,3,4,5\}$ objectives, respectively. In each figure, the horizontal axis indicates the problem parameter α. $\alpha = 1$ is equivalent to DTLZ2, and $\alpha = 100$ is the typical DTLZ4. Also, each figure also shows the 95% confidence intervals.

First, from the results on the problem with $\alpha = 1$ which has the minimum difficulty to obtain uniformly distributed solutions in the objective space, we can see that there is no difference in *HV* values among four algorithms at the final generation. However, *HV* is gradually decreased by increasing α and the difficulty to obtain uniformly distributed solutions. Next, from the results on the problem with $\alpha = 100$, we can see that the proposed MOEA/D-A achieves higher *HV* than the conventional MOEA/D. Also, the proposed MOEA/D-CRU-A achieves higher *HV* than the conventional MOEA/D-CRU. The similar tendency can be seen on DTLZ4 problems with all objectives used in this work. Also, the proposed MOEA/D-CRU-A achieves the highest *HV* on the all problems at the final generation. These results reveal that the proposed chain-reaction solution arrangement contributes to improving the search performance of MOEA/D based algorithms, and its effectiveness becomes significant especially on the problems with a large α which has the difficulty to obtain uniformly distributed solutions in the objective space. However, there is the tendency that the effectiveness of the proposed chain-reaction arrangement is deteriorated on the problems with $\alpha = 1000$. In problems with a large α, since the distribution of randomly generated initial solutions is strongly biased in the objective space, the effect of the proposed method is weakened even if initial solutions are arranged by the proposed method.

5.2 Search Performance over Generations

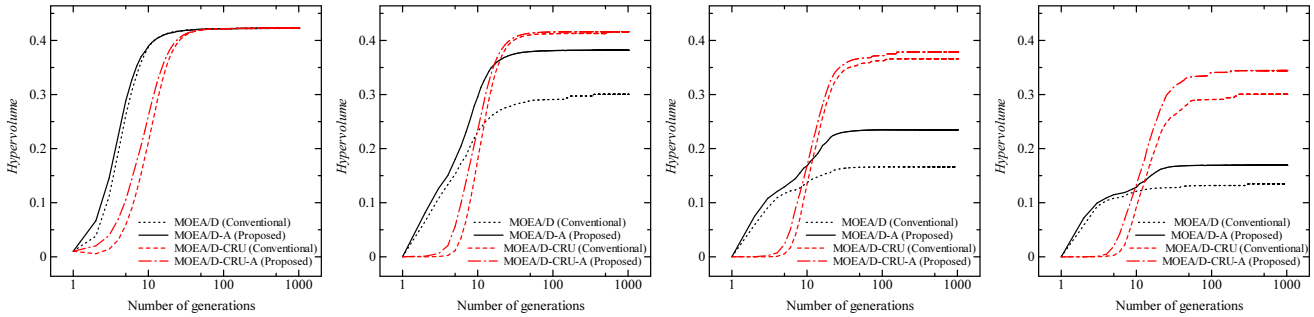
Next, we observe the transitions of the average *HV* values obtained by four algorithms over generations. Figure 3-6 shows the results on problems with 2-5 objectives, respectively. In each of them, (a)-(d) show results on problems with different α, respectively. Note that the horizontal axis is the number of generations and logarithmic scale in all figures.

From the results on problems with $\alpha = 1$ which is equivalent to DLTZ2 with the lowest difficulty, we can see that the proposed MOEA/D-A achieves higher *HV* than the conventional MOEA/D until about 50 generations. The proposed MOEA/D-CRU-A also achieves higher *HV* than the conventional MOEA/D-CRU until about 50 generations. However, after that, the difference of *HV* values of four algorithms disappears. Next, form the results on problems with $\alpha = \{50, 200, 500\}$, in almost all cases, we can see the tendency that *HV* values are saturated after about 100 generations. The proposed MOEA/D-A with the chain-reaction solution arrangement achieves higher *HV* than the conventional MOEA/D without the proposed method, however, both *HV* values are not significantly improved after about 100 generations. On the other hand, until the first 10 generations, *HV* values of MOEA/D-CRU and MOEA/D-CRU-R employing the chain-reaction solution update proposed in our previous work are lower than the ones of MOEA/D and MOEA/D-A without the chain-reaction solution update. However, the algorithms employing the chain-reaction solution update achieve higher *HV* than the algorithm without the one at the final generation. Also, we can see that MOEA/D-CRU-A shows higher *HV* than MOEA/D-CRU throughout the entire search.

These results reveal that the proposed chain-reaction solution arrangement improves the search performance of MOEA/D algorithms at any generation number.

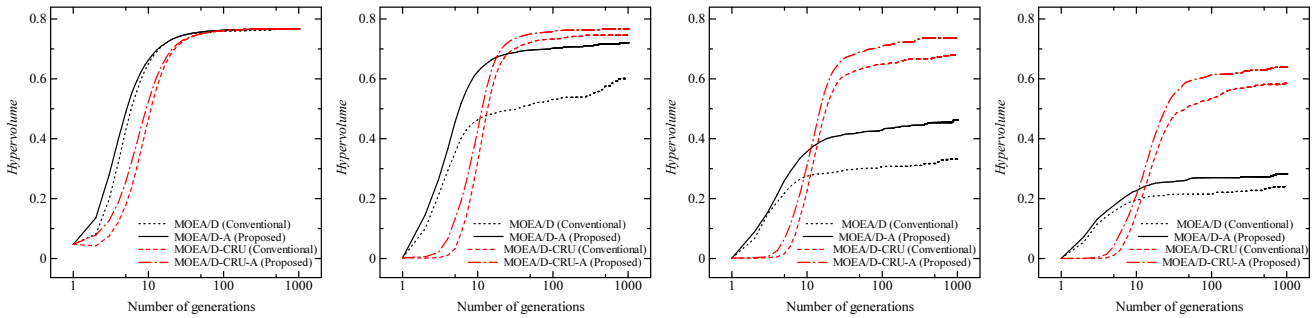
6 Conclusions

To improve the search performance of MOEA/D based algorithms on multi-objective problems by enhancing the simultaneous optimization of many single objective problems directed by weight vectors, this work proposed the chain-reaction solution arrangement method to appropriately arrange randomly generated initial solutions to weight vectors based on their positions in the objective space. The proposed method is an extension of the chain-reaction solution update and calculates the objective balance vector of each initial solutions and tries to assign it to an appropriate weight vector close the objective balance vector. The experimental results using DTLZ4 problems showed that the proposed chain-reaction solution arrangement contributes to improving the search performance of MOEA/D based algorithms.



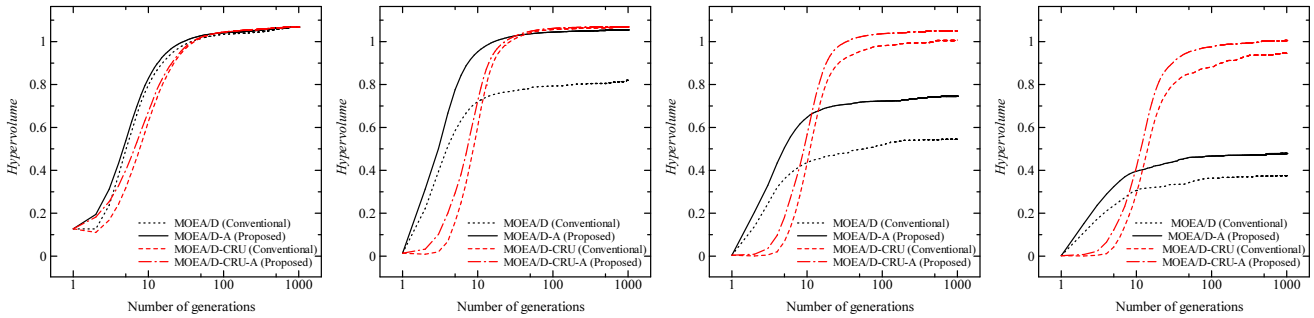
(a) $\alpha = 1$ (DTLZ2) (b) $\alpha = 50$ (c) $\alpha = 200$ (d) $\alpha = 500$

Figure 3. Transitions of HV in DTLZ4 with $m = 2$ objectives and different difficulty parameters α



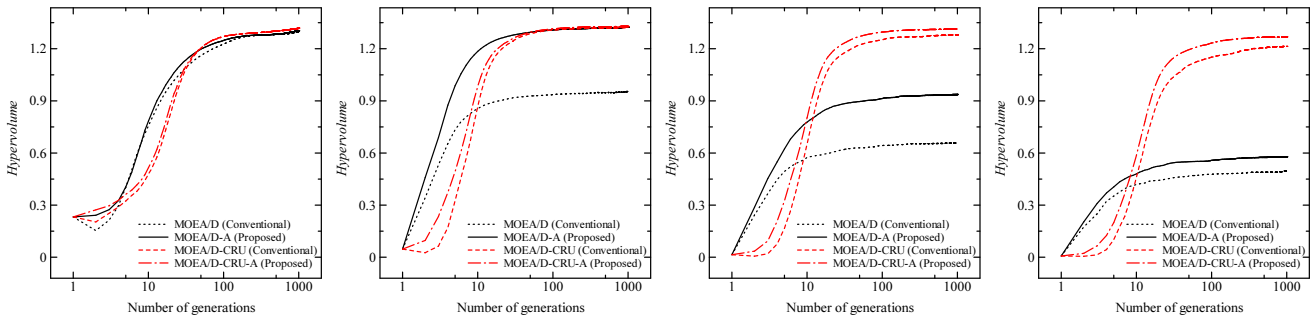
(a) $\alpha = 1$ (DTLZ2) (b) $\alpha = 50$ (c) $\alpha = 200$ (d) $\alpha = 500$

Figure 4. Transitions of HV in DTLZ4 with $m = 3$ objectives and different difficulty parameters α



(a) $\alpha = 1$ (DTLZ2) (b) $\alpha = 50$ (c) $\alpha = 200$ (d) $\alpha = 500$

Figure 5. Transitions of HV in DTLZ4 with $m = 4$ objectives and different difficulty parameters α



(a) $\alpha = 1$ (DTLZ2) (b) $\alpha = 50$ (c) $\alpha = 200$ (d) $\alpha = 500$

Figure 6. Transitions of HV in DTLZ4 with $m = 5$ objectives and different difficulty parameters α

As a future work, we will verify the effects of the proposed method on problems with many objectives and discrete solution spaces.

References

J. Bader and E. Zitzler. HypE: An Algorithm for Fast Hypervolume-based Many-objective Optimization. *Evolutionary Computation, MIT Press*, 19(1):45–76, 2011.

N. Beume, B. Naujoks, and M. Emmerich. SMS-EMOA: Multiobjective Selection Based on Dominated Hypervolume. *European Journal of Operational Research*, 181(3):1653–1669, 2007.

K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*, John Wiley & Sons, 2001.

K. Deb, L. Thiele, M. Laumanns, and E. Zitzler. Scalable Multi-objective Optimization Test Problems. *In Proc. of 2002 IEEE Congress on Evolutionary Computation (CEC2002)*, pages 825–830, 2002.

K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A Fast and Elitist Multi-Objective Genetic Algorithm: NSGA-II. *IEEE Trans. on Evolutionary Computation*, 6(2):182–197, 2002.

K. Deb and H. Jain. An Evolutionary Many-objective Optimization Algorithm Using Reference-point Based Non-dominated Sorting Approach, Part I: Solving Problems with Box Constraints. *IEEE Trans. on Evolutionary Computation*, 18(4): 577–601, 2014.

E. J. Hughes. Evolutionary Many-objective Optimisation: Many Once or One Many? *In Proc. of 2005 IEEE Congress on Evolutionary Computation (CEC'05)*, pp. 222–227, 2005.

K. Li, Q. Zhang, S. Kwong, M. Li, and R. Wang. Stable Matching Based Selection in Evolutionary Multiobjective Optimization. *IEEE Trans. on Evolutionary Computation*, 18(6):1–15, 2014.

H. Sato. Chain-Reaction Solution Update in MOEA/D and Its Effects on Multi and Many-Objective Optimization. *Soft Computing*, Springer, 20(10):3803–3820, 2016.

Q. Zhang and H. Li. MOEA/D: A Multi-objective Evolutionary Algorithm Based on Decomposition. *IEEE Trans. on Evolutionary Computation*, 11(6):712–731, 2007.

E. Zitzler, M. Laumanns, and L. Thiele. *SPEA2: Improving the Strength Pareto Evolutionary Algorithm. TIK-Report*, No.103, 2001.

E. Zitzler and S. Kunzili. Indicator-based Selection in Multiobjective Search. *In Proc. of the 8th Intl. Conf. on Parallel Problem Solving from Nature (PPSN VIII)*, LNCS, Vol. 3242, pages 832–842, 2004.

On Demand Response Modeling and Optimization of Power in a Smart Grid

Olli Kilkki Kai Zenger

Department of Electrical Engineering and Automation, Aalto University School of Electrical Engineering, Finland
{first}.{last}@aalto.fi

Abstract

The electrical grid is under reform; the increasing volatile renewable energy production and distributed local generation compel the development of a future where the consumption of electricity can also participate in maintaining the production-consumption balance of the grid. There is a vast amount of recent research activity related to exploiting residential and industrial consumption elasticity. This paper presents a selected overview of various facets of modeling, optimizing and simulating the demand response potential and effects, especially with emergent soft computing methods. In addition, some illustrating examples are presented, where various relevant approaches from recent state-of-the-art research, including soft computing methods, are reviewed.

Keywords: Smart grid, optimization, soft computing

1 Introduction

The continuing increases in electricity consumption and volatile production in the form of renewable energy generation (Ela et al., 2011), are motivating the renewal of the electrical grid. This coming smart grid is envisioned to be more comprised of electronically controlled equipment than the current electromechanical foundations (Amin and Wollenberg, 2005). All the levels of the grid from the production, through transmission and distribution, to consumption will be fitted with metering and control devices to allow more granular and swift control (Farhangi, 2010).

One main part of the future smart grid is participation of electricity demand in maintaining grid health (Albadi and El-Saadany, 2007). The development of this *demand response* (DR) is motivated by the possible decrease in controllable production, as well as increase in local distributed generation (Castillo-Cagigal et al., 2011). In addition, local energy storages in residential or industrial facilities, such as electric vehicles (Kempton and Tomić, 2005), heat storages (Ericson, 2009) or batteries (Vytelingum et al., 2011), could be utilized for further enhancing the DR capabilities.

Before the actual distribution, electricity is bought and sold on various markets. These include, especially in Europe, some kind of day-ahead auction markets, where a price for the electricity is determined for all periods within the day (Imran and Kockar, 2014). Various intra-day markets for further trading energy are also usually present. In addition, the real-time balance of the production and consumption of electricity is maintained by an assigned system operator (Kirschen and Strbac, 2004). The operator aims to ensure the balance by issuing imbalance fees to producers and consumers who do not adhere to their day-ahead plans, purchasing regulating energy from various market participants and contracting reserves for continuous system balancing.

The demand-side could then be possibly utilized in various stages of market operation (Palensky and Dietrich, 2011). In the first place, smart appliances could enhance energy efficiency. With dynamic electricity prices, the time of use of appliances and other consumption could be shifted to minimize congestion during peak consumption. Then, with more dynamic control, the elasticity of the consumption could be utilized to participate e.g. in intra-day markets to shift consumption for agreed upon compensation. Finally, with more real-time communication and control capabilities, the demand-side could participate in various reserves.

These different options for market participation, in conjunction with the more distributed nature of electricity production and grid health maintenance, provide an ample range of problems and possibilities. The relevant control problems required for demand response participation have to be modeled carefully, due to their dual nature involving the competing objectives related to grid stability and level of comfort (Callaway and Hiskens, 2011). In addition, the amount of uncertainties is going to even increase (Varaiya et al., 2011), and thus has to be taken into account both in the modeling and optimization stages, as well as the final simulations. Various traditional and more novel approaches can be found to offer solutions to the aforementioned problems. Especially soft computing and computational intelligence methods are seen to be able to respond to the various complicated problems at hand (Venayagamoorthy, 2011).

Similarly to previous reviews on the subject of computational intelligence methods in the smart grid context (Venayagamoorthy, 2011), this paper presents the possibilities of soft computing methods, but also with consideration for traditional approaches. This paper presents a selected overview of modeling, optimizing and simulating demand response potential, in addition to some exam-

ples. In the examples, various relevant approaches from recent state-of-the-art research, including soft computing methods, are reviewed, and discussion presented on possible future research directions. At first, we outline the elements and actors to be modeled related to demand-side management. Then we delve into the control of those elements with planning consumptions schedules and then more real-time control related to the grid frequency, while showing some examples. Finally, some discussion on the findings, and challenges and possibilities of the demand response are presented.

2 Modeling demand-side management

There are multiple facets to modeling the effects and potential of demand response in a smart grid. Traditionally electricity delivery and consumption has been a hierarchical system, with electricity provided by large power plants and consumed by customers behind the transmission and distribution networks (Kirschen and Strbac, 2004). However, with the coming smarter electricity grid, the direction of electricity transfer can vary and role of the various actors is expanded (Amin and Wollenberg, 2005). The objectives of these actors and their communication have to be properly modeled, especially if they are to operate in an independent manner. The main players in the electricity market to be modeled include the electricity markets, electricity producers, independent system operator, retailers and finally electricity consumers (Kirschen and Strbac, 2004). In addition, the actual grid, or parts of it, might have to be modeled depending on the effects that are under study.

2.1 Markets

Auction-based electricity markets facilitate the trading of electricity by aggregating production and consumption bids of electricity, and determining an hourly price (Kirschen and Strbac, 2004). Electricity retailers participate on these markets on behalf of the end-consumers by aggregating their consumption. If the customer electricity consumption schedules can be affected, the aggregating retailers can reduce electricity acquisition costs by acquiring more electricity during lower priced hours.

The markets can be modeled by the retailers by assuming that their effect on the market is minimal, and modeling the resulting electricity price as a stochastic process (Zugno et al., 2013). The model for the price can take into account various influences, including climate and weather data, hydro-power availability and electricity demand (Vehviläinen and Pyykkönen, 2005). In addition, the uncertainties in the realizations manifest as normal variation as well as larger price peaks (Voronin et al., 2014). Various methods can be utilized in modeling and forecasting the prices, such as ARMA, GARCH, neural networks and GMMs (Voronin et al., 2014). Alternatively, the price of electricity can be assumed to directly reflect the amount of demand. For example, in many studies the cost of electricity is set to rise quadratically w.r.t. the amount of total consumption (Mohsenian-Rad et al., 2010; Samadi et al., 2010).

2.2 Actors

The aggregating retailer in the context of a smart grid is often referred to as a virtual power plant (VPP), which can aggregate the consumption as well as elasticity of the consumption, and possible distributed production, of multiple industrial or residential customers (Pudjianto et al., 2007). The VPP has to be modeled with both the technical and commercial roles in mind (Pudjianto et al., 2007). When participating in the various markets and providing system balancing services, the VPP has to be able to affect the consumption profiles of its managed customers. There are various alternative approaches for controlling the consumption, which are discussed in more detail in the following Section 3.

Conversely, the consumers are mainly concerned with maintaining their level-of-comfort or production deadlines. In addition, depending on the contract made with the VPP, they have to be able to shift their consumption either under command or voluntarily through compensation (Albadi and El-Saadany, 2007). The customers can have various loads that they use their electricity on. For example, various thermostatically controlled loads such as direct heating or refrigerators could be modeled and utilized for various grid maintenance activities (Callaway and Hiskens, 2011). One general application involves a stochastic load and some intermediate energy storage, such as storage space heating (Ali et al., 2014). Ali *et al.* (Ali and Koivisto, 2013) devised such a model for a particular consumer n, where the storage charging $P_{n,t}$ is constrained by the heating demand $Q_{n,t}$ and storage limits $C_{max,n}$:

$$0 \leq P_{n,t} \leq P_{max,n} \tag{1}$$

$$\sum_{k=1}^{t} \left(P_{n,k} - Q_{n,k}\right)\Delta t \geq -C_{n,0} \tag{2}$$

$$\sum_{k=1}^{t} \left(P_{n,k} - Q_{n,k}\right)\Delta t \leq C_{max,n} - C_{n,0} \tag{3}$$

A more detailed model would be required for industrial applications for providing demand response (Ding et al., 2014), Examples of models for industrial applications include production planning with detailed constraints for operating mode switching with proper ramping behaviour by Mitra *et al.* (Mitra et al., 2012), and the work by Castro *et al.* which takes into account discrete events such as due dates and utility availabilities (Castro et al., 2009).

2.3 Other elements

In addition, other elements have to be taken into account when modeling demand response potential and effects on a larger scale. Locational aspects affect the realization of demand response on a larger scale. Among

these locational aspects are the communication networks enabling coordination of the demand response (Güngör et al., 2011). The communication infrastructure of the smart grid is envisioned to consist of a combination of existing networks and technologies (Güngör et al., 2011), where the latencies and other characteristics are important considerations (Lu et al., 2013). These communication channels can be included in the models by integrating external communication network simulators (Mets et al., 2011), or more directly by utilizing various probability distributions for the relevant parameters (Kilkki et al., 2014). In addition, constraints on locational grid limits might have to be taken into account (Wang et al., 2015).

3 Controlling consumption

Responsive demand can be utilized for various system balancing activities on various timescales (Palensky and Dietrich, 2011). Related to the various services that the VPP can offer to the grid and markets, are the ways in which the control can be exerted on the consumer loads. The main classification of different demand response programs, based on the ways the consumers are incentivized and contracted to react, can be started with a split into incentive based programs and price based programs (Albadi and El-Saadany, 2007). In incentive based programs the incentives can be paid out either by directly assuming them into the consumer contracts or by determining them with on market based methods, e.g. bidding.

In contrast, with price-based programs, the VPP or other controlling entity would choose a dynamic price for the electricity it is offering to its customers (Albadi and El-Saadany, 2007). The type of dynamic price can vary from day-ahead chosen time-of-use prices, to critical peak pricing, or real-time pricing.

3.1 Optimizing scheduling

Two major areas of research involve planning of consumption and charging schedules, and then dispatching the actual consumption. The main objectives usually include taking into account the utility of the consumers, cost of electricity and any possible grid constraints. Traditional optimization methods such as linear programming (LP), mixed-integer linear programming (MILP) or quadratic programming (QP), or a mixture thereof, are routinely utilized in obtaining consumption schedules. Electric vehicle charging aggregation is a thoroughly researched topic, where the charging schedules can be optimized for example w.r.t. uncertain renewable generation (Pantoš, 2011), or taking into account the risks involved in the costs and profits (Momber et al., 2015).

In addition, heating is another large part of electricity consumption which could be temporarily deferred. For example, Nguyen *et al.* (Nguyen and Le, 2014) propose a method for planning heating schedules w.r.t the day-ahead electricity acquisition costs, while considering various uncertainties. In Section 2.2 we detailed a simplified model for an energy storage in conjunction with a heating load, and we have previously also proposed (Kilkki et al., 2015a) an optimization algorithm for its charging schedules. We utilized a genetic algorithm to optimize the dynamic price K_t an aggregator charges from its independent consumers, in order to achieve a desirable aggregate consumption profile. The consumers aim to minimize their electricity costs by optimizing

$$\min_{\text{w.r.t. } P} \sum_{t=1}^{H} K_t P_{n,t} \tag{4}$$

while holding the constraints (1)-(3). The aggregators profits comprise of the expectation function

$$\max \mathbb{E}\Big[\sum_{t=1}^{H} K_t(P_t + L_t) - K_t^s E_t - \pi_t^{\uparrow}\Delta E_t^{\uparrow} + \pi_t^{\downarrow}\Delta E_t^{\downarrow}\Big] \tag{5}$$

where P_t is the aggregate consumption, K_t^s the day-ahead market price, $\pi_t^{\uparrow/\downarrow}$ and $\Delta E_t^{\uparrow/\downarrow}$ the imbalances, costs and payments for positive and negative deviations from the planned schedules, respectively. The optimization aims to minimize these costs as well as the acquisition costs, while maximizing profit that can be acquired from the consumers. The imbalances that are however realized during the day, were proposed to be minimized by offering favorable changes to the electricity price of the consumers for them to shift their consumption. The price discount amount and percentage of consumers to offer to were also optimized using genetic algorithms with the to-be-maximized function defined as

$$\max_{\text{w.r.t. } \Delta K,n} \pi_t^{\uparrow/\downarrow}(\Delta E_t^{\uparrow/\downarrow} + P_t^d - P_t) + \sum_{k=t}^{H} n(K_k^d P_k^d - K_k P_k) \tag{6}$$

where K_k^d is the discounted price (with discounts $\Delta K \leq 0$) and P_k^d the consumption of the consumers after the discount. Figure 1 shows results obtained from the aforementioned optimizations, when discounts were given at hours 3, 14 and 20. The upper subfigure shows the realized load consisting of the storage space heater charging and other inflexible consumption sources. The lower subfigure them shows the distribution of realized imbalances from the acquired electricity schedule, that occur due to various uncertainties. It can be seen how during the hours with discounts, the amount of imbalances is reduced significantly, resulting in positive effects for the aggregator, consumers, and the grid.

An evolutionary approach has also been utilized for various other problems, such as by Logenthiran *et al.* for scheduling general shiftable loads (Logenthiran et al., 2012). In addition to genetic algorithms, various other soft computing methods have been utilized in planning consumption schedules (Venayagamoorthy, 2011). For example, Soares *et al.* developed a day-ahead scheduling method for electric vehicles, with considerations for demand response. The resulting mixed-integer nonlinear programming problem was solved utilizing traditional

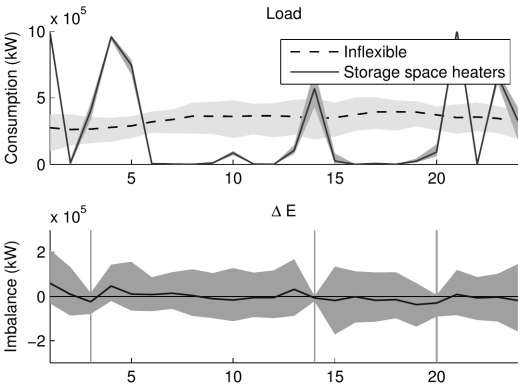


Figure 1. Realized charging and inflexible consumption as well as remaining imbalances from the acquired schedules after discounts. (Kilkki et al., 2015a)

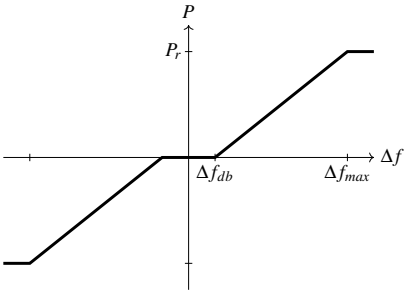


Figure 2. Droop.

nonlinear optimization methods but was found to be significantly faster to solve with a particle swarm optimization approach (Soares et al., 2013). Various fuzzy logic (Dubey et al., 2015; Ma and Mohammed, 2014) and neural network solutions (Siano et al., 2012), and combinations thereof (Ozturk et al., 2013; Shahgoshtasbi and Jamshid, 2011) have been also employed, in optimizing day-ahead electricity consumption schedules.

3.2 Controlling (via) frequency

The instantaneous balance of total aggregate production and consumption in the electrical grid can be inferred to some extent from the frequency of the grid (Kundur et al., 2004). The system operator is in charge of maintaining this frequency close to its nominal value. The maintenance is performed by purchasing regulating power to minimize larger imbalances, and contracting either power plants or consumption for providing continuous reserves. Involving the consumption in this frequency control is widely researched with applications to power plants such as hydro (Doolla and Bhatti, 2006) and wind turbines (Ramtharan et al., 2007), as well as to demand.

In its most simplest form, frequency regulation can be implemented by implementing the traditional droop control with consumers increasing and decreasing their loads w.r.t. to the deviation of grid frequency from its nominal value (Palensky and Dietrich, 2011). Figure 2 displays the traditional droop of a frequency controller, where P_r is the amount of reserves promised, Δf_{db} is the deadband of the frequency deviations and Δf_{max} the frequency deviation where the maximal control has to be exerted. Thermostatical loads can be similarly co-ordinated to perform the droop control by issuing frequency limits after which they turn their loads on or off, respectively.

Molina-García *et al.* (Molina-Garcia et al., 2011) proposed a similar method while including time characteristics to the frequency control thresholds. In their proposed approach, different load types were given different characteristics on how fast to respond to varying sizes of deviations of varying lengths. For an alternative approach, Callaway proposed an aggregated continuous control method for a group of thermostatically controlled loads (Callaway, 2009), which could be utilized for frequency regulation. The control was achieved by directly altering the temperature setpoints of the thermostats, while maintaining an aggregated load model of all of the loads under control. Callaway and Hiskens also proposed similar methods for energy storage charging devices, such as electric vehicles, where a hysteresis-based on-off charging cycle is manipulated by adjusting the deadbands of the cycles (Callaway and Hiskens, 2011).

In addition to the more traditional approaches, the various proposed methods have utilized soft computing solutions. A fuzzy logic based frequency control method was developed by Datta (Datta, 2014), where electric vehicle charging was controlled to provide frequency regulation. Similar fuzzy methods were also utilized for photovoltaics inverter control (Datta and Senjyu, 2013). Bevrani & Shokoohi utilized a fuzzy approach as well in their work (Bevrani and Shokoohi, 2013), where they developed a model-free generalized droop control scheme for microgrids. The model-free controller was obtained by a neural network approach, where the generated network was trained with historical training data to achieve the desired co-ordinated droop control. Debbarma and Dutta propose (Debbarma and Dutta, 2016) a fractional order controller for controlling the charging of electric vehicles. A flower pollination algorithm was then utilized for tuning the actual parameters of the controller. The flower pollination algorithm is a metaheuristic optimization algorithm (Yang, 2012), which takes inspiration from the pollination process of flowers with elements such as seperation to local and global pollination and a random walk based on the Lévy flight.

There have also been multiple studies on taking into account frequency regulation capacity in day-ahead planning and market operation. For example, the author has previously proposed an approach for scheduling the charging of energy storages utilized in heating, while including participation in frequency reserve markets (Kilkki et al., 2015b). Conversely, Yuen *et al.* proposed an algorithm for provisioning reserves from multiple microgrids (Yuen et al., 2011). Vayá and Andersson developed a method for planning the day-ahead charging schedules of an aggrega-

tor, while simultaneously optimizing for frequency regulation participation (Vayá and Andersson, 2013), as well as a real-time dispatch algorithm for the regulation. The aggregated vehicle charging patterns utilized in the optimization of the day-ahead schedules were obtained by utilizing a co-evolutionary algorithm for optimization. Similarly optimizing electric vehicle charging was also proposed by Sortomme and El-Sharkawi (Sortomme and El-Sharkawi, 2012). Their algorithm also takes into account reserve participation, and additionally includes in the optimization the uncertainties in departure times of the vehicles.

4 Discussion and conclusions

The modernification of the electrical grid brings with it many challenges. As the grid transforms from a more hierarchical system to a more diverse collection of more dynamic actors, the control and its various effects have to be carefully considered. However, many opportunities also arise as the penetration of relevant ICT devices can enable responsive demand. Both residential and industrial demand could be included in various demand-side management programs, ranging from day-ahead scheduling and market bidding, to more real-time applications such as participation in frequency reserves. For modeling and optimizing the responsiveness of the demand, various methods can be utilized. Traditional methods can provide a more predictable response, but with the increase in the complexity of the system, various soft computing methods become attractive alternatives. We found multiple proposed applications of genetic algorithms and fuzzy methods in scheduling the consumption, as well as in coordinating frequency control. Soft computing methods we found to be utilizable in the optimization of scheduling, as well as tuning of more real-time control. In addition, with the various uncertainties involved in the optimization and control of consumption, various methods are required for identifying the time-series models related to the uncertainties. Computational intelligence methods are again often utilized.

References

MH Albadi and EF El-Saadany. Demand response in electricity markets: An overview. In *IEEE Power Engineering Society General Meeting*, volume 2007, pages 1–5, 2007.

Mubbashir Ali and Matti Koivisto. Optimizing the DR control of electric storage space heating using LP approach. *International Review on Modelling and Simulations (IREMOS)*, 6(3): 853–860, 2013.

Mubbashir Ali, Juha Jokisalo, Kai Siren, and Matti Lehtonen. Combining the demand response of direct electric space heating and partial thermal storage using LP optimization. *Electric Power Systems Research*, 106:160–167, 2014.

S Massoud Amin and Bruce F Wollenberg. Toward a smart grid: power delivery for the 21st century. *Power and energy Magazine, IEEE*, 3(5):34–41, 2005.

Hassan Bevrani and Shoresh Shokoohi. An intelligent droop control for simultaneous voltage and frequency regulation in islanded microgrids. *Smart Grid, IEEE Transactions on*, 4 (3):1505–1513, 2013.

Duncan S Callaway. Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy. *Energy Conversion and Management*, 50(5):1389–1400, 2009.

Duncan S Callaway and Ian A Hiskens. Achieving controllability of electric loads. *Proceedings of the IEEE*, 99(1):184–199, 2011.

Manuel Castillo-Cagigal, Estefanía Caamaño-Martín, Eduardo Matallanas, Daniel Masa-Bote, A Gutiérrez, F Monasterio-Huelin, and J Jiménez-Leube. PV self-consumption optimization with storage and active DSM for the residential sector. *Solar Energy*, 85(9):2338–2348, 2011.

Pedro M Castro, Iiro Harjunkoski, and Ignacio E Grossmann. New continuous-time scheduling formulation for continuous plants under variable electricity cost. *Industrial & engineering chemistry research*, 48(14):6701–6714, 2009.

Manoj Datta. Fuzzy logic based frequency control by V2G aggregators. In *Power Electronics for Distributed Generation Systems (PEDG), 2014 IEEE 5th International Symposium on*, pages 1–8. IEEE, 2014.

Manoj Datta and Tomonobu Senjyu. Fuzzy control of distributed PV inverters/energy storage systems/electric vehicles for frequency regulation in a large power system. *Smart Grid, IEEE Transactions on*, 4(1):479–488, 2013.

Sanjoy Debbarma and Arunima Dutta. Utilizing electric vehicles for LFC in restructured power systems using fractional order controller. *Smart Grid, IEEE Transactions on*, 2016.

Yue Min Ding, Seung Ho Hong, and Xiao Hui Li. A demand response energy management scheme for industrial facilities in smart grid. *Industrial Informatics, IEEE Transactions on*, 10(4):2257–2269, 2014.

Suryanarayana Doolla and TS Bhatti. Load frequency control of an isolated small-hydro power plant with reduced dump load. *Power Systems, IEEE Transactions on*, 21(4):1912–1919, 2006.

Hari Mohan Dubey, Manjaree Pandit, and BK Panigrahi. Hybrid flower pollination algorithm with time-varying fuzzy selection mechanism for wind integrated multi-objective dynamic economic dispatch. *Renewable Energy*, 83:188–202, 2015.

Erik Ela, Michael Milligan, and Brendan Kirby. Operating reserves and variable generation. *Contract*, 303:275–3000, 2011.

Torgeir Ericson. Direct load control of residential water heaters. *Energy Policy*, 37(9):3502–3512, 2009.

Hassan Farhangi. The path of the smart grid. *Power and energy magazine, IEEE*, 8(1):18–28, 2010.

Vehbi C Güngör, Dilan Sahin, Taskin Kocak, Salih Ergüt, Concettina Buccella, Carlo Cecati, and Gerhard P Hancke. Smart grid technologies: communication technologies and standards. *Industrial informatics, IEEE transactions on*, 7(4): 529–539, 2011.

Kashif Imran and Ivana Kockar. A technical comparison of wholesale electricity markets in North America and Europe. *Electric Power Systems Research*, 108:59–67, 2014.

Willett Kempton and Jasna Tomić. Vehicle-to-grid power implementation: From stabilizing the grid to supporting large-scale renewable energy. *Journal of Power Sources*, 144(1): 280–294, 2005.

Olli Kilkki, A Kangasrääsiö, Raimo Nikkilä, Antti Alahäivälä, and Ilkka Seilonen. Agent-based modeling and simulation of a smart grid: A case study of communication effects on frequency control. *Engineering Applications of Artificial Intelligence*, 33:91–98, 2014.

Olli Kilkki, Antti Alahaivala, and Ilkka Seilonen. Optimized control of price-based demand response with electric storage space heating. *Industrial Informatics, IEEE Transactions on*, 11(1):281–288, 2015a.

Olli Kilkki, Christian Giovanelli, Ilkka Seilonen, and Valeriy Vyatkin. Optimization of decentralized energy storage flexibility for frequency reserves. In *Industrial Electronics Society, IECON 2015-41st Annual Conference of the IEEE*, pages 002219–002224. IEEE, 2015b.

Daniel Kirschen and Goran Strbac. *Fundamentals of Power System Economics*. Wiley Online Library, 2004.

Prabha Kundur, John Paserba, Venkat Ajjarapu, Göran Andersson, Anjan Bose, Claudio Canizares, Nikos Hatziargyriou, David Hill, Alex Stankovic, Carson Taylor, et al. Definition and classification of power system stability ieee/cigre joint task force on stability terms and definitions. *Power Systems, IEEE Transactions on*, 19(3):1387–1401, 2004.

Thillainathan Logenthiran, Dipti Srinivasan, and Tan Zong Shun. Demand side management in smart grid using heuristic optimization. *Smart Grid, IEEE Transactions on*, 3(3): 1244–1252, 2012.

Xiang Lu, Wenye Wang, and Jianfeng Ma. An empirical study of communication infrastructures towards the smart grid: Design, implementation, and evaluation. *Smart Grid, IEEE Transactions on*, 4(1):170–183, 2013.

Tan Ma and Osama A Mohammed. Optimal charging of plug-in electric vehicles for a car-park infrastructure. *Industry Applications, IEEE Transactions on*, 50(4):2323–2330, 2014.

Kevin Mets, Tom Verschueren, Chris Develder, Tine L Vandoorn, and Lieven Vandevelde. Integrated simulation of power and communication networks for smart grid applications. In *Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2011 IEEE 16th International Workshop on*, pages 61–65. IEEE, 2011.

Sumit Mitra, Ignacio E Grossmann, Jose M Pinto, and Nikhil Arora. Optimal production planning under time-sensitive electricity prices for continuous power-intensive processes. *Computers & Chemical Engineering*, 38:171–184, 2012.

Amir-Hamed Mohsenian-Rad, Vincent WS Wong, Juri Jatskevich, Robert Schober, and Alberto Leon-Garcia. Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid. *Smart Grid, IEEE Transactions on*, 1(3):320–331, 2010.

Angel Molina-Garcia, François Bouffard, and Daniel S Kirschen. Decentralized demand-side contribution to primary frequency control. *Power Systems, IEEE Transactions on*, 26 (1):411–419, 2011.

Ilan Momber, Afzal Siddiqui, Tomas Gomez San Roman, and Lennart Soder. Risk averse scheduling by a pev aggregator under uncertainty. *Power Systems, IEEE Transactions on*, 30 (2):882–891, 2015.

Duong Tung Nguyen and Long Bao Le. Optimal bidding strategy for microgrids considering renewable energy and building thermal dynamics. *Smart Grid, IEEE Transactions on*, 5 (4):1608–1620, 2014.

Yusuf Ozturk, Datchanamoorthy Senthilkumar, Sudhakar Kumar, and Gene Lee. An intelligent home energy management system to improve demand response. *Smart Grid, IEEE Transactions on*, 4(2):694–701, 2013.

Peter Palensky and Dietmar Dietrich. Demand side management: Demand response, intelligent energy systems, and smart loads. *Industrial Informatics, IEEE Transactions on*, 7(3):381–388, 2011.

Miloš Pantoš. Stochastic optimal charging of electric-drive vehicles with renewable energy. *Energy*, 36(11):6567–6576, 2011.

Danny Pudjianto, Charlotte Ramsay, and Goran Strbac. Virtual power plant and system integration of distributed energy resources. *Renewable power generation, IET*, 1(1):10–16, 2007.

G Ramtharan, Janaka Bandara Ekanayake, and Nicholas Jenkins. Frequency support from doubly fed induction generator wind turbines. *Renewable Power Generation, IET*, 1(1):3–9, 2007.

Pedram Samadi, Amir-Hamed Mohsenian-Rad, Robert Schober, Vincent WS Wong, and Juri Jatskevich. Optimal real-time pricing algorithm based on utility maximization for smart grid. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 415–420. IEEE, 2010.

Dariush Shahgoshtasbi and Mo Jamshid. Energy efficiency in a smart house with an intelligent neuro-fuzzy lookup table. In *System of Systems Engineering (SoSE), 2011 6th International Conference on*, pages 288–292. IEEE, 2011.

Pierluigi Siano, Carlo Cecati, Hao Yu, and Janusz Kolbusz. Real time operation of smart grids via FCN networks and optimal power flow. *Industrial Informatics, IEEE Transactions on*, 8 (4):944–952, 2012.

Joao Soares, Hugo Morais, Tiago Sousa, Zita Vale, and Pedro Faria. Day-ahead resource scheduling including demand response for electric vehicles. *Smart Grid, IEEE Transactions on*, 4(1):596–605, 2013.

Eric Sortomme and Mohamed A El-Sharkawi. Optimal scheduling of vehicle-to-grid energy and ancillary services. *Smart Grid, IEEE Transactions on*, 3(1):351–359, 2012.

Pravin P Varaiya, Felix F Wu, and Janusz W Bialek. Smart operation of smart grid: Risk-limiting dispatch. *Proceedings of the IEEE*, 99(1):40–57, 2011.

M González Vayá and Göran Andersson. Combined smart-charging and frequency regulation for fleets of plug-in electric vehicles. In *IEEE Power and Energy Society General Meeting*, volume 2013, 2013.

Iivo Vehviläinen and Tuomas Pyykkönen. Stochastic factor model for electricity spot price - ťhe case of the nordic market. *Energy Economics*, 27(2):351–367, 2005.

Ganesh Kumar Venayagamoorthy. Dynamic, stochastic, computational, and scalable technologies for smart grids. *Computational Intelligence Magazine, IEEE*, 6(3):22–35, 2011.

Sergey Voronin, Jarmo Partanen, and Tuomo Kauranne. A hybrid electricity price forecasting model for the nordic electricity spot market. *International Transactions on Electrical Energy Systems*, 24(5):736–760, 2014.

Perukrishnen Vytelingum, Thomas D Voice, Sarvapali D Ramchurn, Alex Rogers, and Nicholas R Jennings. Theoretical and practical foundations of large-scale agent-based micro-storage in the smart grid. *Journal of Artificial Intelligence Research*, 42(1):765–813, 2011.

Kun Wang, Zhiyou Ouyang, Rahul Krishnan, Lei Shu, and Lei He. A game theory-based energy management system using price elasticity for smart grids. *Industrial Informatics, IEEE Transactions on*, 11(6):1607–1616, 2015.

Xin-She Yang. Flower pollination algorithm for global optimization. In *Unconventional computation and natural computation*, pages 240–249. Springer, 2012.

Cherry Yuen, Alexandre Oudalov, and Adrian Timbus. The provision of frequency control reserves from multiple microgrids. *Industrial Electronics, IEEE Transactions on*, 58(1): 173–183, 2011.

Marco Zugno, Juan Miguel Morales, Pierre Pinson, and Henrik Madsen. A bilevel model for electricity retailers' participation in a demand response market environment. *Energy Economics*, 36:182–197, 2013.

Application of Musical Expression Generation System to Learning Support of Musical Representation

Mio Suzuki

Department of Creative Engineering National Institute of Technology, Kushiro College 2-32-1 Otanoshike-Nishi, Kushiro, Hokkaido 084-0916, Japan, mio@kushiro-ct.ac.jp

Abstract

This paper proposes a learning support system of musical representation by piano using teacher's example of musical expression that is generated based on impression expressed by an adjective with our musical expression generation system. The system evaluates learner's performance comparing with teacher's example using a Kansei space and fuzzy rules expressing the relationship between musical expression and impression. The system presents good points of learner's performance and advice by text to a learner for improvement of learner's musical representation. A learner tries to improve his/her own performance based on system's advice, and the system presents other advice again. From the experimental results, it is show that the proposed system is useful to learn musical representation and an approach of the proposed system is suitable because the affirmative evaluation is obtained from the participants who have taken piano lesson. On the other hand, it is found that to learn musical representation is difficult using the proposed system for learner of low-performance skills.

Keywords: musical expression, impression, fuzzy inference, learning support system

1 Introduction

Musical expression is the deviation of performance from tempo marks and/or dynamic marking in a score, which means a suitable change of tempo or volume as music in a real performance. Musical expression is varied according to performance situations and/or performers. The same musical piece gives listeners different impressions depending on performances or performers. Then, it is said that performing a score correctly does not necessarily present rich musical expression. We propose a musical expression generation system called MUSAI (MUSical expression generation system by Adjective with Interaction) (Suzuki and Onisawa, 2015) that generate a musical expression reflecting any impression expressed by an adjective, where musical expression is performed by a piano. MUSAI has the relationship between musical expression and impression as knowledge, which is expressed using a Kansei space and fuzzy rules, and generates musical expression reflecting impression expressed by an adjective. Furthermore, if a user does not feel that generated musical expression reflects impression well, MUSAI modifies musical expression based on user's evaluation by the interaction with a user. This paper applies to MUSAI to learning support of musical representation for a piano lesson.

There are some studies on learning support of music as the application of musical performance generation systems (Oshima et al., 2004; Ferrari et al., 2006). These researches make an effect to encourage willingness of a learner to practice because a computer automatically corrects a pitch of musical performance of a learner. If a teacher is not great with pianos, a teacher can show a musical representation without any concerns for a mistake about musical performance. This paper considers the case in which a teacher instructs a learner in musical representation using musical performance and verbal advice, and proposes the learning support system of musical representation with the application of MUSAI.

2 Components and Parameters of Musical Expression

This paper covers the musical expression by a piano. According to (Schmitz, 1977), changes of tempo, volume and length of a note have an influence on the impression of musical performance. Furthermore, a player can change them during the performance of a piece of music. Therefore, in this paper, tempo, volume and length of a note are considered as components of musical expression that are represented by a parameters as shown Table 1.

In this paper, the melody of a piano piece is played with the right hand and the chords are played with left hand. And the musical expression is added to a phrase of piano piece of music.

3 Learning Support System of Musical Representation

The target of the proposed learning support system of musical representation is a piano learner. Figure 1 shows the outline of learning support system of musical representation. The learning support system presents a musical expression as an example to a learner, which is generated by MUSAI. The system also presents impression to a learner using by an adjective. A learner listens to an example of musical expression and practices using a digital piano. A musical expression of a learner is recorded with

Table 1. Parameters of musical expression

Component	Parameter	Value	Meaning
Tempo	*TempoBase*	[40, 208]	Basic tempo of a performance
	TempoRange	[0.0, 0.6]	Variation range of a tempo
	TempoVar	*constant*, *decrease*, *increase*	Change type of a tempo
Volume	*VelocityBase*	[16, 127]	Basic volume of a performance
	VelocityRange	[0.0, 0.7]	Variation range of a volume
	VelocityVar	*constant*, *decrease*, *increase*	Change type of a volume
Length of note	*LengthSign*	-1, 0, 1	Show perform a note for longer or shorter
	LengthBase	[1.5, 10.0]	Show change of note length from in score
	LengthRange	0	Variation range of a length of note
	LengthVar	*constant*	Change type of a length of a note

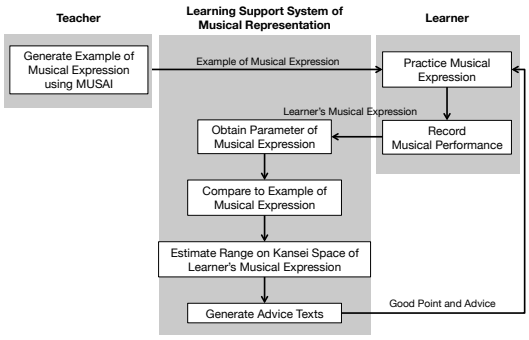


Figure 1. Outline of learning support system of musical representation

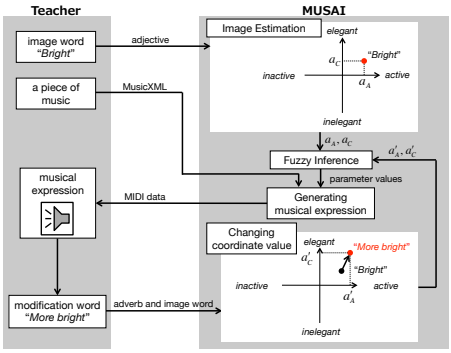


Figure 2. Detail of example of musical expression generation part

MIDI data. The learning support system obtains the parameters values of musical expression from data and compares them with the parameter values of a musical expression example. The system estimates the range on a Kansei space corresponding to learner's musical expression, where a Kansei space is constructed beforehand. The system presents a good point of learner's musical expression and an advice for progress in musical representation based on the comparison and the estimation. A learner continues to practice playing the piano by reference to the presented example of musical expression, the good point and an advice given by the system.

3.1 Generation of Example of Musical Expression

Figure 2 shows the outline of generating an example of musical expression using by MUSAI. A teacher inputs a piano piece as an original musical piece and an adjective expressing musical expression impression, which is called an image word in this paper.

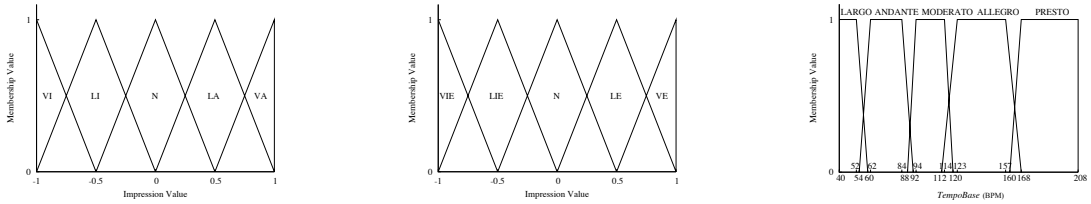
The image word is mapped onto a Kansei space in the sense that the coordinates values on this Kansei space are obtained through the image estimation process in Figure 2, using the concept of co-occurrence of adjectives(Shimizu and Hagiwara, 2006). The parameter values of musical expression are obtained by the coordinates values on a Kansei space and fuzzy inference. MIDI data is generated using the obtained parameter values and an inputted piano piece of music, and then, the generated musical expression is presented to a teacher. If a teacher is not satisfied with the presented musical expression, a teacher inputs a modification word consisting of an adverb and an image word, e.g., *more bright*, which expresses teacher's evaluation of the presented musical expression. The modification part estimates new coordinates values of musical expression on the Kansei space according to the modification word, and new parameter values of musical expression are obtained by fuzzy inference. Then, the presented musical expression is modified. A teacher evaluates the modified musical expression whether he/she is satisfied with it or not. These procedures are repeated until satisfactory musical expression is obtained.

3.1.1 Construction of Kansei Space

A Kansei space is constructed using data obtained by preliminary experiments. In the preliminary experiments, various musical expressions are generated automatically by setting the parameter values of musical expression at random and are presented to the experiment participants. The participants evaluate their impressions of presented musical expressions using the semantic differential method (Osgood et al., 1957) with a 5-points scale. Then, *an active factor* and *an elegant factor* are extracted by factor analysis of evaluation data. A two-dimensional Kansei space consists of these two factors axes.

3.1.2 Image Estimation

An inputted image word is mapped onto the Kansei space in the sense that the coordinates values on this Kansei space are obtained through the image estimation process.



(a) Active factor **(b)** Elegant factor **(c)** *TempoBase*

Figure 3. Membership functions correspond to the premise in fuzzy rules

In this process, at first, the co-occurrence phrase of adjectives is prepared using the inputted image word and pairs of adjectives included in the factors composing the Kansei space (Shimizu and Hagiwara, 2006).

The co-occurrence phrase of adjectives is searched using a Web search and the number of Web pages having the co-occurrence phrase is counted. And the similarity degrees of an inputted image word and the adjectives belonging to the factor of the Kansei space are obtained using the number of Web pages having co-occurrence phrase of adjectives. Coordinates values on the Kansei space are obtained using the similarity degree of an image word and adjectives included in the factors composing the Kansei space.

3.1.3 Parameter Values Estimation using Fuzzy Inference

Parameter values of musical expression are obtained using fuzzy inference by Mamdani's min-max-gravity method (Kruse et al., 1994) from the coordinates values obtained by the image estimation process. In this paper, the following fuzzy rule form is used; if the coordinate value of an image word on the *active factor* axis is $\tilde{A}$ and that on the *elegant factor* axis is $\tilde{B}$, then *TempoBase* is $\tilde{C}$, where $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$ are fuzzy sets. Figure 3(a) and 3(b) show membership functions of fuzzy sets of the premise in fuzzy rules. These are related to the coordinates values on Kansei space obtained by the image estimation. Figure 3(c) shows membership functions of *TempoBase* of the consequent in fuzzy rules. The membership functions of *TempoBase* are defined based on tempo mark corresponding to BPM (Beats Per Minute). Table 2 shows constructed fuzzy rules to estimate a value of *TempoBase*. In this paper, twenty five fuzzy rules are constructed per one parameter value. The weights of the fuzzy rules are same for all fuzzy rules. Refer to (Suzuki and Onisawa, 2015) about other fuzzy sets of the consequent in the fuzzy rules and other fuzzy rules.

3.1.4 Interactive Modification of Musical Expression

Generated musical expression is not necessarily satisfied with a teacher. Therefore, the generated musical expression is modified according to teacher's evaluation and a modification word, e.g., *more bright*. The musical expression is modified by changing the coordinates values on the Kansei space according to teacher's evaluation, and its algorithm is based on Interactive Particle Swarm Optimization (hereinafter, referred to as IPSO) (Madar et al., 2005).

Table 2. Fuzzy rules to estimate value of *TempoBase*

	VIE	LIE	N	LE	VE
VI	Largo	Andante	Andante	Andante	Andante
LI	Andante	Andante	Andante	Andante	Moderato
N	Moderato	Moderato	Moderato	Moderato	Allegro
LA	Allegro	Allegro	Allegro	Allegro	Allegro
VA	Allegro	Allegro	Allegro	Presto	Presto

In IPSO, an individual, i.e., musical expression in this paper, has the position and the velocity, which are calculated by the following equations:

$$\boldsymbol{x}_j(t+1) = \boldsymbol{x}_j(t) + \boldsymbol{v}_j(t), \tag{1}$$

and

$$\begin{aligned} \boldsymbol{v}_j(t+1) &= w(t)\boldsymbol{v}_j(t) + c_1 r_1(\boldsymbol{p}_j(t) - \boldsymbol{x}_j(t)) \\ &+ c_2 r_2(\boldsymbol{g}(t) - \boldsymbol{x}_j(t)) \end{aligned} \tag{2}$$

where $\boldsymbol{x}_j(t)$ is the position of j-th individual at t-th iteration, $\boldsymbol{v}_j(t)$ is the velocity of j-th individual at t-th iteration, $r_1, r_2 \in [0,1]$ are uniformly distributed random numbers, $\boldsymbol{p}_j(t)$ and $\boldsymbol{g}(t)$ are the best position of j-th individual and the best position of population at t-th iteration, respectively, w is defined by (3), and c_1' and c_2' defined by (4) and (5) are transformed to c_1 and c_2.

$$w(t) = (RangeHigh_j(t) - RangeLow_j(t)) \times adv, \tag{3}$$

$$c_2' = \frac{1}{1 + e^{-(EvalP_j - 5)}}, \tag{4}$$

and

$$c_1' = 1.0 - c_2' \,. \tag{5}$$

In this paper, $RangeHigh_j(t)$ and $RangeLow_j(t)$, search ranges are defined as $+1.0$ and -1.0 at the first modification step, and the search range becomes narrow at every

iteration, and *adv* is defined according to a modification word as follows: 0.25 when *a little more*, 0.5 when *more* and 0.75 when *very*. As for c_1 and c_2, c_1' and c_2' obtained by (4) and (5) are transformed by adding 0.5 to the values in [0.5, 1.5]. $EvalP_j$ is the number of musical expressions that a teacher's evaluation is affirmative for j-th individual, i.e., j-th musical expression. According to $EvalP_j$, c_1 and c_2 are changed based on (4) and (5) so that musical expression at $(t+1)$-th iteration moves to the best position of j-th musical expression or the best position of population at t-th iteration.

Musical expressions are generated at the first modification step as follows. The Kansei space is divided into 25 ranges according to the fuzzy division as shown in Figure 3(a) and 3(b). The position, i.e., the coordinates values of modified musical expression are moved to some range according to the adverb of the modification word as follows: the next range when *a little more*, the next range but one when *more* and the next range but two when *very*. At this time, the coordinates values are determined at random in the range, and the velocity of modified musical expression is defined as zero.

Modified musical expressions are presented to a teacher again and a teacher evaluates whether the modified musical expression reflects impression expressed by an image word or not as follows:

-1: not reflecting,
0: neutral,
$+1$: rather reflecting,
$+2$: reflecting.

And a teacher chooses one musical expression reflecting an image word best out of presented musical expressions as the best one. When a teacher wants to modify musical expression again, a teacher inputs a modification word. The velocity and the position of each musical expression are updated by (1) and (2) according to the evaluation value for musical expression at the previous step and the inputted modification word. After the second modification step, the number of modified musical expression is set as 10. If the number of musical expressions is smaller than 10, musical expressions are generated at random so that the number is 10. If there are ten or more musical expressions at and after the second modification step, top ten musical expressions having high evaluation survive at the next modification step. Above procedures are repeated until a teacher is satisfied with generated musical expressions. In this way, the MUSAI obtains an example of musical expression of a musical work for practice by a teacher.

3.2 Comparison of Parameter of Musical Expression

The presented learning support system obtains the parameter values of learner's musical expression, compares its parameter values with those of an example of musical expression, and evaluates whether a learner plays a piano according to an example of musical expression presented by the system. The eight parameter values chosen from ones shown in Table 1, excluding *LengthBase* and *LengthRange*, are considered for the comparison. Obtained parameter values of learner's musical expression are evaluated from the following point of views: whether or not fuzzy sets which the parameter values of learner's musical expression belong to are the ones which the same parameter values of an example of teacher's musical expression belong to, where the fuzzy sets are the ones in the consequent part of fuzzy rules mentioned in Section 3.1. Based on the comparison results the learning support system presents some good points of learner's musical expression and some advices for improvement.

3.2.1 Advice for Parameter of Musical Expression

If some parameter value of learner's musical expression belongs to the fuzzy set which the same parameter value of teacher's example of musical expression belongs to, the following advice is presented to a learner: "You play the performance well for the parameter."

If all parameter values of learner's musical expression belong to the same fuzzy sets which the parameter values of teacher's example of musical expression belong to, the text is generated as good point the following and presented to a learner: "Your musical expression reflects impression expressed by an image word."

On the other hand, if the parameter values of learner's musical expression does not belong to the fuzzy set which the same parameter value of a teacher's example of musical expression belong to, advice text is presented to a learner showing how to play a performance.

3.3 Range of Learner's Musical Expression on Kansei Space

Learner's performance of musical expression is mapped to a range on the Kansei space using its parameter values. Text sentences are generated using the coordinate values of teacher's example of musical expression, those of learner's one and information on the practice effect. Text sentences present that what extent learner's musical representation reflects target impression and that what extent impression of learner's musical expression is changed by playing performance of musical expression according to the presented advice.

3.3.1 Estimation of Range of Musical Expression

The mapped range of learner's musical expression on the Kansei space is estimated using the parameter values of learner's musical expression and the fuzzy rules in MUSAI in the following way. At first, as for the parameter value of learner's musical expression that does not belong to the same fuzzy sets which the same parameter values of teacher's musical expression belongs to, the ranges of the active factor and the elegant factor axes composing the Kansei space are estimated from the fuzzy rules having

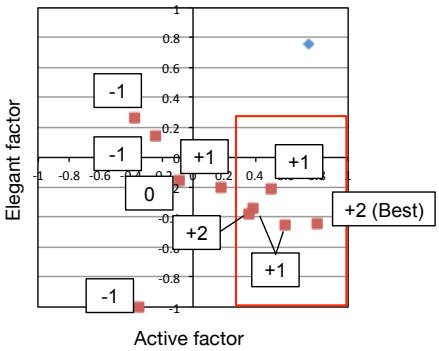


Figure 4. Example of evaluation values in estimated range

the fuzzy sets in the consequent part which the parameter value of learner's musical expression belongs to. Next, the overlapped range of the obtained range on the active factor axis and that on the elegant factor axis is obtained on the Kansei space.

3.3.2 Advice about Impression

The system verifies whether or not the coordinate values of teacher's example of musical expression are included in the overlapped range on the Kansei space. If the coordinate values are included in the range, advice about impression of learner's musical expression is presented to a learner using the evaluation values of teacher's example of musical expressions included in the range, the number of musical expression example with the same evaluation, and inputted modification words.

The system estimates whether target impression is reflected by learner's musical expression or not using evaluation values of playing performances of learner's musical expression. Let teacher's example of musical expression get evaluation values during generation of teacher's example as shown in Figure 4. Impression of learner's musical expression is estimated as *rather* as shown below because the number of evaluation value $+1$ is the largest.

> Your musical performance has **rather bright** impression.

If any teacher's example of musical expression are not included in the overlapped range, the system does not give any advice.

The system gives a learner another advice using the center coordinate values of the obtained range and the coordinate values of a teacher's example of musical expression on the Kansei space. According to the distance between the center coordinate values of the range and the coordinate values of a teacher's example of musical expression, i.e., $dist$, the degree adverb in advice is selected as follows: *a little more* when $0 \leq dist \leq 0.25$, *more* when $0.25 < dist \leq 0.5$, and *very* when $0.5 < dist \leq 0.75$. As for the image word in advice, the system chooses the image word according to the number of affirmative expressions in the modification words and the number of negative ones in the modification words during practice. For instance, let us assume that the system gives a learner affirmative advice *more bright*, and *a little more bright*, and negative advice, *not bright* during the practice. In this instance, a learner is given affirmative advice twice and negative advice once. Then, the image word *bright* is chosen as advice. If the musical expression is not included in the estimated range or the number of affirmation advices is equal to or less than the number of negative ones, the following text is presented to a learner.

> Play the piano as given advice, then your performance of musical expression has **bright** impression.

4 Experiment

Experiment is performed in order to verify the usefulness of the proposed learning support system of musical representation. Participants are 10 males or females of 16 through 26 years old. 8 out of all participants have taken a piano lesson and other 2 participants have never taken it. In the experiment, two examples of musical expression are generated by an experimenter using MUSAI. That is, an experimenter is a piano teacher. And the participants have practice in musical representation using a digital piano and proposed learning support system until the system evaluates that learner's musical expression reflects impression, where the upper limit number of practice repetition is fixed at 10.

Table 3 shows questionnaire items in this experiment. The participants answer the questionnaire items from Q1 to Q4 every one practice, and answer Q5 and Q6 after the experiment with the 7-points scale. Points from 1 to 3 mean negative evaluation and points from 5 to 7 mean affirmative evaluation, and point 4 means neutral. Musical pieces used in this experiment are Gnossienne No. 1 written by Erik Satie and Ecossaise in G Major, WoO.23 written by Ludwig van Beethoven. The example of musical expression of Gnossienne No. 1 reflects hard impression and the example of musical expression of Ecossaise in G Major, WoO.23 reflects light impression.

Table 3. Questionnaire items

	Item
Q1	Is advice helpful?
Q2	Do you think the contents of advice is suitable?
Q3	Do you have a easy to understand advice?
Q4	Do you think you can perform musical expression close to an example by advice text?
Q5	Do you think you will learn musical representation using this system?
Q6	Do you think you want to practice musical representation using this system?

4.1 Results and Discussions

Figure 5 shows the evaluation results. Although all participants musical expressions are not evaluated by the system that they reflect target impression, it is found that from Figure 5 that affirmation ratios in Q1, Q2 and Q4 are 90.0%, 95.0% and 88.2%, respectively. This means that the participants feel that presented advices are useful for the practice of musical representation using the learning support system. It is also found from Figure 5 that the affirmation ratios in Q5 and Q6 are 80.0% and 90.0%, respectively. This means that the participants feel that the learning support system is useful to learn musical representation. The participants have taken a piano lesson who have the following free descriptions about advice presented by the learning support system: "The system is useful because the system evaluates my performance objectively when practicing the piano alone." and "It was easily practice because the system presents good points and advice for my performance." Furthermore, the two participants have never taken a piano lessen who have the following free descriptions: "I think good about the system presents quantified my performance." and "It was easy to understand that attend to my performance because the system presents example of musical expression and advice."

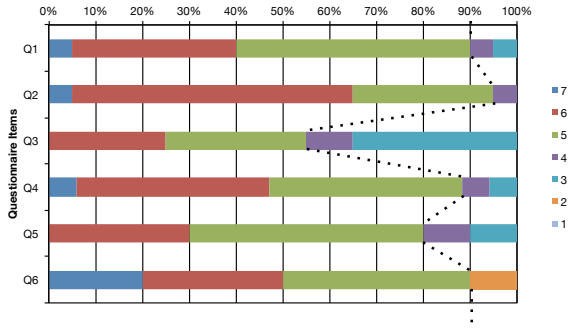


Figure 5. Evaluation results for learning support system

On the other hand, the affirmation ratio in Q3 is 50.0% lower than its in other questionnaire items. It is found the participants feel that it is hard to understand advice presented by the learning support system. Some participants have free descriptions about the easiness of understanding advice: "I understand the presented advice, but I don't imagine how I should improve my performance." One of the reasons is that although the system compares learner's performance with a teacher's example of musical expression, it does not compare the current learner's performance with learner's former ones.

From the above experimental results, it is said that proposed learning support system is useful to learn musical representation. And an approach of learning support of musical representation by the proposed system is suitable because the affirmative evaluation is obtained from the participants who have taken a piano lesson.

Figure 6 show the changes of the average number of

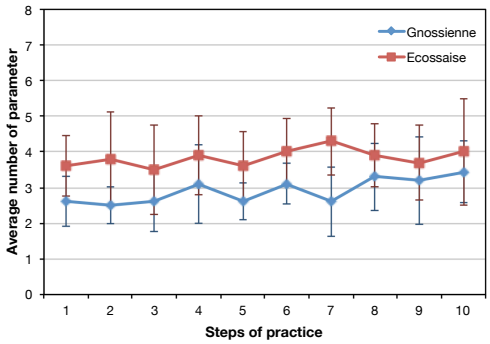


Figure 6. Change of average number of parameters of Gnossienne's and Ecossaise's musical expression

parameter values of participant's musical expression for Gnossienne and Ecossaise matching to the parameter values of teacher's example in the sense that fuzzy sets in the consequent part of fuzzy rules mentioned in Section 3.1 which the parameter values of participant's musical expression belong to are the same as the ones which the parameter values of teacher's musical expression belong to. Although there is no significant difference between the matching degree at the first practice and that at the tenth practice for the both performances, the increasing tendency is observed from these figures. In fact, the correlation coefficient between the number of times of practice and the average value of the matching degree is 0.771 for Gnossienne, and 0.496 for Ecossaise. This means that the parameter values of participant's musical expression approach to those of teacher's example of musical expression as the participants repeat practice. On the other hand, in the participant H and I who have never taken a piano lesson, the correlation coefficient between the number of times of practice and the matching degree of participant H for Gnossienne and Ecossaise are -0.572 and -0.078 respectively, and that of participant I for Gnossienne and Eccosaise are 0.232 and -0.087 respectively. One reason for this is the piano performance skill of these 2 participants is not high.

From the above, although to learn the musical representation is difficult for a learner having low performance skills by learning support system, it is said that a learner can learns how to play performance of musical expression by learning support system's advice.

5 Conclusions

This paper proposes the learning support system of musical representation using the musical expression generation system called MUSAI and verifies the validity of the proposed learning support system by the experiment. The learning support system obtains the parameter values of musical expression from learner's performance. And the system compares the parameter values of learner's musical performance with those of an example of musical expression and gives a learner an advice about good points and improvement of learner's musical expression. From

the experimental results, it is found that proposed learning support system is useful to learn the musical representation. And it is said that the approach of proposed learning support system is suitable because the affirmative evaluation is obtained by the participants who have taken a piano lesson. On the other hand, it is found that to learn the musical representation is difficult for a learner of low performance skills using proposed system. A performance skills of a learner is considered as future works.

References

L. Ferrari, Anna R. Addessi, and F. Pachet. New technologies for new music education: The continuator in a classroom setting. In *Proceedings of ICMPC 06*, 2006.

R. Kruse, J. Gebhardt, and F. Klawonn. *Foundations of FUZZY SYSTEMS*. JOHN WILEY & SONS, England, 1994.

J. Madar, J. Abonyi, and F. Szeifert. Interactive Particle Swarm Optimization. In *Proceedings of 5th International Conference on Intelligent Systems Design and Application*, pages 314–319, 2005.

Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. *The measurement of meaning*. University of Illinois Press, 1957.

C. Oshima, K. Nishimoto, and M. Suzuki. Family Ensemble: A Collaborative Musical Edutainment System for Children and Parents. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, pages 556–563, 2004.

Hans-Peter Schmitz. *Singen und Spielen : Versuch einer allgemeinen Musizierkunde*. Symphonia, 1977. S. Imoto and K. Takii, trans.

K. Shimizu, and M. Hagiwara. Image Estimation of Words Based on Adjective Co-occurrences. *The IEICE transactions on information and systems*, 89(11), pages 2483–2490, 2006.

M. Suzuki, and T. Onisawa. Musical expression generation reflecting user's impression by interaction. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, 27(4), pages 651–668, 2015.

Verifying an Implementation of Genetic Algorithm on FPGA-SoC using SystemVerilog

Hayder Al-Hakeem Suvi Karhu Jarmo T. Alander

Department of Electrical and Energy Engineering, University of Vaasa, Finland, {firstname.lastname}@uva.fi

Abstract

In this paper we show how an efficient implementation of genetic algorithms can be done on Field Programmable Gate Array i.e. on programmable hardware using the latest hardware design language aiding verification. A fourway number partitioning problem of 128 unsigned 16-bit integers is used as a test case of the implementation. However, other similar problems could be solved using the proposed approach. The design was implemented using a combination of reusable verified intellectual property cores for arithmetic operations and VHDL to describe the genetic algorithm operators in register transfer level. The register transfer level components were verified in ModelSim using SystemVerilog assertions and covergroups. Test results show significant improvements in performance compared to C language implementation running on a core i-7 desktop computer.

Keywords: genetic algorithms, verification, FPGA, system on chip (SoC)

1 Introduction

The idea of using hardware to speed up processing of evolutionary algorithms is not new (Alander, 2008; Alander et al., 1995). Here we show how the implementation can be done in a way that uses the latest verification techniques that are available in modern hardware design languages such as SystemVerilog.

The proposed Genetic Algorithm (GA) implementation is tested on a number partitioning problem that belongs to the set of NP complete problems meaning that its solution might be intractable in practice. However, many real life optimization problems belong to the NP complete set and must be somehow solved approximately for practical purposes. Having enough computing power helps somewhat. Field Programmable Gate Arrays (FPGA) are energy efficient and fast due to their massive parallel processing. In problems that they are suitable, they can be significantly faster than a PC while using only a fraction of the energy of a corresponding processor solving the same task. An obvious drawback of FPGAs is that they need both programming and hardware design skills. Thus, creating high quality implementation solution on FPGA is both demanding and needs a lot of testing and verification. Prototype Verification System (Owre et al., 1992) has been used to verify crossover operator in GAs (Nawaz et al., 2013). And vice versa, GAs have also been used to optimize verification (Gao et al., 2015; Cheng and Lim, 2014).

2 Formulating the GA Optimization

The basic principle of GA is that if some randomly generated solutions can produce good results, those solutions can be combined and used as building blocks to generate better solutions. Solutions are evaluated by calculating a fitness function, they are then modified using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. The new generated solutions are then re-evaluated and the procedure is repeated until the target of fitness optimization is achieved or a predefined number of iterations is reached. (Alander, 1992)

2.1 The Problem Encoding

The number partitioning problem is often labeled as the easiest hard problem (Hayes, 2002). While being considered as one of the classical NP-hard problems of combinatorial optimization, it is fairly easy to understand, represent, and evaluate.

According to (Korf, 2009), "The number partitioning problem is to divide a given set of integers into a collection of subsets, so that the sum of the numbers in each subset are as nearly equal as possible". In this work, a four-way partitioning problem is used to verify the functionality and benchmark the performance of our hardware GA implementation. The goal is to split a set of N=128 randomly generated positive integers I_i of length 16 bits into four subsets so that the sums of those subsets are as equal as possible.

Figure 1 illustrates the chromosome (solution trial) representation.

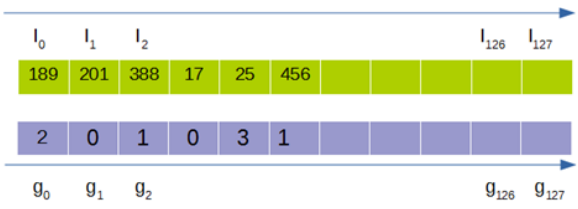


Figure 1. Chromosome representation when N=128

Each chromosome consists of N genes. A binary representation is used, where each gene consist of two bits. Each gene g_i (i=0, .., N-1) can take one of the following

four binary values; "00", "01", "10" or "11" indicating that the corresponding integer I_i belongs either to subset 0, 1, 2 or 3 respectively. A brute force search would require the evaluation of $4^{128} = 1.158 \times 10^{77}$ possible solutions which is intractable.

2.2 The Objective Function (Fitness)

For n number of subsets, the summation of each subset (S_0, S_1, S_2,..., S_n) is computed first. Matlab simulations showed that the standard deviation or variance produced good results. However, to reduce the complexity and cost of hardware implementation, a more simplistic fitness function is implemented. The function adds the difference between every possible permutation pair of the subset sums as illustrated in Equation 1. The objective is to minimize the fitness.

$$\sum_{i=0}^{n-1} \sum_{j=i+1}^{n} |(S_i - S_j)| \quad (1)$$

In case of a 4-way partitioning problem (n=4), unpacking equation 1 yields to Equation 2.

$$Fitness = D_{0,1} + D_{0,2} + D_{0,3} + D_{1,2} + D_{1,3} + D_{2,3} \quad (2)$$

Where $D_{i,j}$ is the absolute value of the difference between S_i and S_j

2.3 Selecting The GA Operators

Since the GA is intended to be implemented and verified in FPGA, the simplest possible set of operators that can provide satisfactory results were investigated using Matlab simulations. In these simulations, the number of generations was fixed at 2000. For each set of operators, the experiment was repeated 1000 times. For each experiment, the error was defined as the difference between the largest and smallest of the four subsets sums divided by the sum of the whole set. If the average error of the 1000 experiment is $\leq 1\%$, this insures that the sum of each of the four subsets is deviating less than 1% from the optimal 25% of the whole set sum. The operators are considered good enough and a simpler set of operators is simulated. After running several simulations, the simplest set of operators that was able to achieve an average error less than 1% (0.23%) was chosen as follows:

- Tournament Selection was used.
- One point crossover was used.
- For Mutation, a positive random integer x is generated for each gene, where $0 \leq x \leq a$, a is a constant that is used to adjust the mutation rate. The random integer x is then compared to a predefined constant integer c, where $0 \leq c \leq a$. If x equals c, the corresponding gene is replaced with two random bits. Otherwise no change is performed. Therefore, the mutation rate can be controlled by changing a, and is defined by $MR = 1/(a+1)$.
- For replacement, the chromosome with worst (largest) fitness value is replaced with the new offspring before the generations counter is incremented.

2.4 Tuning the GA Parameters

Matlab simulations are used to fine-tune the population size, mutation rate, number of generations and the tournament size. Each time, one parameter is varied and the rest are set to fixed values. One thousand experiments, each with a new set of 128 random integers are performed and the best fitness of each experiment is recorded. After 1000 experiments have been performed, the results of those experiments are averaged. The standard deviation is also taken into consideration in order to guarantee a lower probability of getting a bad solution which is highly deviated from the average result.

To evaluate the best size of the selection tournament, The number of generations was set to G=2000, the mutation rate was MR= 1%, and the population size was set to $N_p = 128$. Table 1 shows the averaged best fitness of 1000 experiments and its standard deviation using different tournament sizes. The best results are obtained when the tournament size is 32/128 (i.e. 1/4 of the population size N_p). Since 16 bit numbers produce large sums and therefore large fitness values, the fitness values are illustrated with the k (kilo) metric prefix for convenience.

Table 1. Optimizing tournament size $\mathbf{N_T}$.

$\mathbf{N_T}$	Average Fitness	STD of Fitness
4/128	12.187k	7.068k
8/128	7.201k	4.103k
16/128	7.057k	4.310k
32/128	**6.623k**	**4.158k**
64/128	6.8646k	4.575k

To evaluate the best mutation rate, the number of generations was set to G=2000, the $N_p = 128$ and the tournament size was set to $N_T = \frac{1}{4}N_p = 32$. Table 2 shows the averaged best fitness of 1000 experiments and its standard deviation using different mutation rates. Best results are obtained when MR=1%.

Table 2. Optimizing mutation rate **MR**.

MR	Average Fitness	STD of Fitness
0.25%	10.924k	6.448k
0.5%	8.518k	5.513k
1%	**6.623k**	**4.158k**
2%	6.686k	3.944k
4%	13.656k	6.909k

To evaluate the best population size, The following parameters were fixed; G=2000, the MR=1% and $N_T = \frac{1}{4}N_p$. Larger populations did not provide significant improvements as seen from table 3.

Table 3. Optimizing population size $\mathbf{N_P}$.

N_P	Average Fitness	STD of Fitness
16	**7.026k**	**4.246k**
32	6.892k	4.369k
64	7.085k	4.219k
128	6.623k	4.158k

As for the number of generations, no significant improvement was observed after G=2000. Hence, a fixed number of 2000 generations is selected to simplify hardware implementation.

Based on the simulation results, the following parameters were selected for implementation; N_p=16, N_T=4, MR=1% and G=2000.

3 FPGA Implementation

The simulated GA was implemented using RTL (Register Transfer Level) description in VHDL. Each of the GA operators was implemented in a separate VHDL file to simplify the verification process. A top module called GA instantiates, connects and controls the operation of the GA operators using a finite state machine. Arithmetic operations were implemented using Altera's verified fixed point numbers IP cores. The design is optimized to achieve the best possible performance at the expense of size.

3.1 GA's HDL Entities

3.1.1 Fitness

In order to evaluate the sum of integers in each subset, a separate parallel adder with N=128 inputs is instantiated for each subset. The integers are multiplied with a binary mask before being inputted to the parallel adder to zero the integers that do not belong to the subset in question. This approach is illustrated in Figure 2.

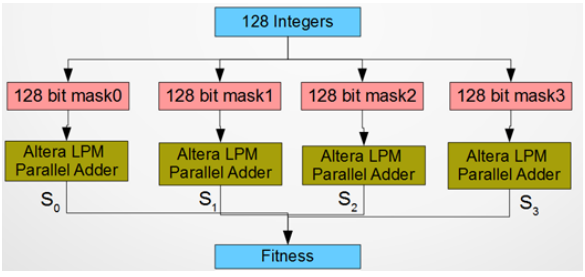


Figure 2. Fitness Circuit

The bit masks are generated from the genes of the chromosome being evaluated. For example, if the fourth gene g_4 in the chromosome has the binary value "11" (decimal three), this means that the integer with index 4 belongs to subset 3. Consequently, bit 4 in mask3 will be assigned '1' while bit 4 in the three other masks will be assigned '0' as shown in Figure 3.

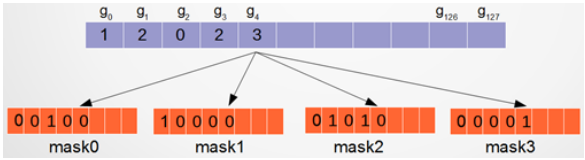


Figure 3. Adder's Masks

Fitness is then calculated from the subset sums using equation 2. The fitness operator is thus purely combinatorial requiring only a single clock cycle to sample and store the result.

3.1.2 Selection

The selection entity accepts 4 random candidates and uses a pair of comparators to output the best two candidates in a single clock cycle.

3.1.3 Crossover

The crossover entity has two parents and a 7-bit random integer (crossover point index) as input. Each clock cycle, it copies half of the genes (higher than the crossover point index) from parent 1 and the other half of genes from parent 2 and outputs a new offspring.

3.1.4 Mutation

For each new offspring, a positive random integer x is generated, where $0 \leq x \leq N-1$. N is the number of integers to be partitioned which is equal to the number of genes in each chromosome. The gene that has an index equal to x is replaced with a new randomly generated gene $\in [0,3]$. As a result, the mutation rate MR=1/128 = 0.78125% which is close to the best mutation rate of 1% obtained from Matlab simulations.

3.1.5 Generating Pseudo Random Sequences

In order to generate random numbers, maximal sequence linear feedback shift registers (LFSRs) are used. Tap locations are obtained from Xilinx tables (Alfke,1996). A separate LFSR is used for each required random number and initialized with a different seed to improve randomness.

The population LFSR is a 256-bit LFSR used to generate the initial population and replace mutated genes with new random values. The selection LFSR is 16 bits wide where each four bits are interpreted as an index from 0 to 15 allowing the selection entity to randomly pick 4 parent candidates each clock cycle. The crossover LFSR is 7 bits wide and it describes the single point crossover location. Similarly, the mutation operator requires a 7-bit long index for selecting which gene will be mutated.

3.1.6 Replacement

In order to identify the index of the individual with the worst fitness, a 16 numbers comparator consisting of 4 stages tree of two number comparators is implemented. The output of each stage comparators named "a greater than or equal b" (ageb) is used to backtrack the tree and identify the index of the individual with the worst (largest) fitness. Figure 4 illustrates the process using, for convenience, a smaller 8 number 3 level comparator. Each level produces 1 bit of the index of the largest fitness.

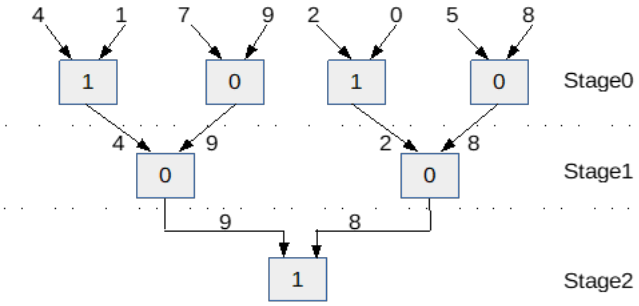


Figure 4. Largest fitness evaluation for 8 values using 3 levels tree

Each of the boxes represents a comparator of two integers. The number inside the box represents the binary result of $A>B$ comparison where A is the left hand side input integer and B is right hand side input integer. Backtracking from stage 2 to stage 0 we obtain the binary sequence "100" which indicates that the largest number exists in index 4 (indexing from right to left). The same procedure can be scaled to compare a larger number of integers adding one extra stage of comparators and one extra bit to the index's LSB each time the number of compared integers is doubled. However, resource utilization and timing constrains must be carefully examined.

3.2 Performance Evaluation

A SystemVerilog *testbench* was created to simulate the proposed GA in Modelsim. A state named report is added to the state machine to be executed after reaching the maximum generation to send the results before resetting the GA.

The *testbench* uses a class that generates N random positive integers from a user defined seed and stores them in an array. Those random integers are buffered serially to the GA. The GA then initializes the population and runs for 2000 generations. After that, the GA buffers thirteen 32-bit words to the *testbench*. The first 8 words are the best solution (chromosome) with the most significant word buffered first. The next 5 words are the best fitness, S_0, S_1, S_2, S_3 and finally a counter of the number of clocks since the GA has started receiving the input.

The SystemVerilog *testbench* then uses the reported best solution to split the integers into 4 subsets and calculate the sum of each subset. The fitness value is calculated using equation 2. The *testbench* will then assert if the calculated sums and fitness match the ones reported by the GA implemented in VHDL. In case of a mismatch, an error message is printed and simulation is interrupted. Otherwise, results are printed and a new set of random integers is generated and buffered to the GA.

Initializing the integers requires 128 clock cycles (in case there are no wasted clock cycles by the input between successive integers). N_p number of clock cycles are required to initialize the population, one item per cycle. In each generation, 4 clock cycles are required to sequentially calculate selection, crossover, mutation, and finally the replacement of the worst individual with the new offspring. Additional N_p clock cycles are required to compare the fitness values of the final generation and extract the best solution. Thirteen clock cycles are consumed reporting the results. Finally, 4 extra clock cycles are required for switching between other internal states in the state machine. The total number of required clock cycles thus becomes (equation 3).

$$CLK(G,N_p) = 4G + 2N_p + 145 \tag{3}$$

ModelSim simulations report a clock count of 8177 when G=2000 generations and N_p=16, which matches exactly Equation 3. Using a clock with 25MHz, the required time per experiment is 8177/25MHz = 0.32708 milliseconds. In other words, the expected performance of the implemented GA is 3057 experiments (each consisting of partitioning N=128 integers) per second.

3.3 Functional Verification

The verification is performed with SystemVerilog assertions and covergroups. We use random inputs and require that certain coverage for each input will be achieved. At each iteration step we check that the results of the operator under verification are correct. Measuring the coverage is an essential step in verification (Wile et al., 2005). In addition to verifying the results, we verify that certain intermediate results inside the modules are correct. For this purpose we need to create new output ports to the modules. The verification is performed by a person other than the designer, but the approach is still that of grey-box verification where the verifier is aware of certain features of the source code of the system.

3.3.1 Fitness

In the verification of the fitness module we verify that the sums and the final fitness value are correct. Since the amount of possible values for the input chromosome is large (2256), we divide the possible values into 216 bins and require that each of them will be covered. In addition, we require that the four special cases where all the genes of the chromosome are the same (000000_2, 010101_2, 101010_2 or 111111_2) will be covered. We also require that at least 6 of the cases where all the numbers belong only to 2 groups will be covered, one for each combination of groups. The same is required for 3 groups, where 4 cases are required, one for each combination of

groups. Also a case where all 4 groups are present must be covered. The numbers to be partitioned are 16-bit and every possible value for them must be covered. However, since there are 128 numbers, not each value for every number have to be covered, but each value must occur at least once among the numbers. In addition we require that the case where all the numbers to be partitioned have the maximum value (216), will be covered. This case must also be cross covered with the cases where all the numbers belong only to 1, 2 or 3 groups. In this way we verify that no overflow occurs in the fitness calculation.

In the calculation of the sums the fitness module utilizes masks that tell which group each number belongs to. Each number should belong to exactly one group. To verify this we sum the masks pointwise. The sum should be 1 at each index. In this test we use the same coverage requirements as when verifying the results.

3.3.2 Selection

The selection module selects the parents for the next crossover. We verify that the operator selects the correct parents specified by the input *selection LFSR* and the fitness values, and reports the correct worst individual. There are sixteen 25-bit fitness inputs, and for each of them we create 216 bins which must be covered. Every value of the 16-bit *selection LFSR* must be covered. In addition we require that at least one of the cases where all the fitness values are the same will be covered.

3.3.3 Replacement

The selection module also contains the replacement functionality. We verify that the replacement works as intended. For this purpose we add a new port which contains the array of fitness values that are stored inside the module. We require that each possible value for the new fitness input will be covered. We also verify that the selection works correctly after the replacement. For that purpose we use the same coverage requirements as described in the Selection section.

3.3.4 Crossover

The crossover module is a rather simple, one-point crossover. The crosspoint is a 7-bit input ranging from 0 to 127. We require that every possible value of the 7-bit crosspoint will be covered. We especially focus on the case where the value is 0 or 127. The crossover component should change the value 0 to 1 and the value 127 to 126, because otherwise the result would comprise only of parent1 or parent2. For both 256-bit parents we create 216 bins which must all be covered. Because the area near the crossing point is the most error-prone, we require that per each index, every combination of 6 gene values around the index is covered, 3 genes for both parents. We require that a case where the parents are the same is covered.

3.3.5 Mutation

The inputs of the mutation module are the 256-bit child to be mutated, the 7-bit index to the gene to be mutated and the random 256-bit individual, the *population LFSR*, which will provide the new content for the gene to be mutated. We verify that the operator will mutate the correct gene with correct value specified by the inputs. We require that for the child and the *population LFSR*, 216 bins will be covered. We also require that all 27 values for the index will be covered. No cross coverage requirements will be set.

Each gene of the child has a probability of P=1/128 to proceed into mutation phase, where the content of the gene will be replaced with the content from the corresponding gene from the *population LFSR*. We call the probability P as the internal mutation rate. The probability that the new content is different than in the original gene is C=3/4. Thus the overall probability that the content of a particular gene will change is $CP = C \times P = 3/512$. We call C as the external mutation rate and CP as the overall mutation rate.

We verify that the mutation rate of the component is correct. We use the same coverage requirements as when verifying the results. We examine one gene and check if the content will change in CP of the cases. We repeat this for each gene. If the actual mutation rate differs from the correct rate, it may be because the *testbench* does not generate purely random but pseudo random values for the index and *population LFSR* inputs. To verify that this is the reason we make a test where the effect of the external mutation rate is eliminated, by keeping the *population LFSR* always different than the child to be mutated. Now the result should be close to P. If the result still differs, it may be due to the fact that the index input is also pseudo random. However the difference should be smaller than when the effect of the external mutation rate is present.

3.3.6 LFSRs

The sequence generated by an LFSR is a binary numeral system just like natural binary code or Gray code. Eventually the shift register enters a repeating cycle. To obtain good pseudo random numbers the cycle should be as long as possible This is called maximal length sequence. The maximal cycle length is $2^n - 1$ where n is the number of bits in the LFSR.

We wanted to verify that the sequences generated by the LFSR module were maximal-length. For the shortest outputs this is possible by using covergroups. During a cycle of $2^n - 1$ executions every possible value except all ones should be covered. With the 256-bit population LFSR we cannot do this. For this output we will simply verify that the sequence is correct for the first 216 cases. We can do this because we know the seed and the random sequence generating function. We verify the correctness of the sequence also in the case of the smaller LFSRs. At least 216 first numbers are checked for each output sequence.

3.3.7 Comparator

The system also contains a comparator module for determining the largest of sixteen 25-bit numbers. For each input we create 216 bins which must be covered. Also

at least one of the cases where the numbers are the same must be covered.

4 Integrating GA into the SoC

In this work, GA was implemented and tested on Terasic's SoCKit hardware, which contains the 925 MHz, Dual-Core ARM Cortex-A9 MPCore Hard Processor System (Terasic, 2015).

4.1 Interfacing GA with the HPS

In order to communicate with and control the GA from the HPS, Xillybus IP core was used. As the authors describe it in their website, "An FPGA IP core for easy DMA over PCIe with Windows and Linux" (Xillybus, 2017). Xillybus comes with a Linux distribution called Xillinux that runs on the FPGA embedded ARM processor and communicates with the Xillybus IP core via a device driver. The driver allows the developers to easily communicate with the FPGA using C language's *read()* and *write()* commands with FIFO buffers. Xillybus allows users to create online accounts to customize then download IP cores. The IP core factory is available to try for free for researchers (Xillybus, 2017). However, it requires licensing for commercial usage. Using Xillybus provides several advantages:

- Streaming data to and from FPGA is done using DMA without impacting the performance of the operating system.
- Xillybus supports a wide range of both Altera and Xilinx FPGAs and is part of the official Linux Kernel device drivers making the design more portable.
- The same design can also be used by connecting the FPGA as a coprocessor via *PCI Express* with an external host running either Microsoft windows or Linux.

Xillybus bus clock runs at 100 MHz supporting a maximum bandwidth of 400 M-Bytes/sec (32-bit interface). To analyze timing requirements for the GA clock, Altera's Time Quest Timing Analyzer was used. A timing netlist based on the "Slow 1100mV 85C Model" with 0 IC delays option unchecked was used to reflect the worst possible case scenario. Altera's tool reported an F_{max} of 36.99 MHz. The GA was tested practically on the device and found to operate correctly at 50 MHz. However, the GA clock was restricted to 25 MHz to ensure reliable operation in the worst conditions. This clock is derived from the Xillybus bus clock using Altera's Phased Locked Loop IP core. In order to enable cross clock domain communication between GA and Xillybus, two Altera *DCFIFO* (dual clock FIFO buffers) entities were instantiated and connected to the GA inside a new top module. The result top module has only standard FIFO ports plus the required clocks and resets and can be directly port mapped to any HDL design that contains standard 32-bit FIFO buffers without any further modification.

4.2 Controlling GA from the Linux host

The Xillinux is a standard Ubuntu desktop for ARM processor that comes with the gcc compiler and a rich preinstalled collection of packages and libraries.

In order to test the GA on the SoCKit, a C program that performs the same functionalities as the SystemVerilog *testbench* was written and compiled using gcc running on Xillinux. The C program opens Xillybus drivers which exist under the Linux /dev directory using the C *open()* command. It then uses the C write() command to write random integers to the Xillybus *write* FIFO buffer and waits for the GA report. When GA is finished, it buffers the results to Xillybus *read* FIFO buffer and resets waiting for new integers. The results are then read by the C program using the C *read()* command and verified.

If no errors are detected, the C program visualizes the results by using the Linux system API to call *GNUplot* and pass the sums of the initial partitions as well as the optimized sums reported by the FPGA. *GNUplot* generates and displays two charts stacked horizontally to visually compare the optimized and non-optimized partitions.

On the SoC FPGA, the reported clock count averaged at 200000 $\pm$ few thousand clock cycles per experiment. Running 1000 experiments required about 10 seconds. However, this is much slower than the reported clock counts by Modelsim's simulations. After examining the buffers interfaces using Altera's Signal Tap 2 Logic analyzer, it was found that the C function *write()* is causing large delays. A quick remedy is to increase the size of the FIFO buffers and write all the required integer for 1000 experiments (1000*128*4=512000 bytes) as one block. This indeed reduced the delay of performing 1000 experiments to less than 2 seconds.

For performance comparison, the same GA was implemented using C language and compiled using *GNU GCC* "gcc (tdm64-1) 5.1.0" with *CodeBlocks* on a desktop computer running Windows 10. The PC has has an Intel core i-7 4770 @ 3.4 GHz with 8 GB DDR3. Ten experiments were perfomed. In each experiment, 1000 number partitoning problems are solved. The average time was calculated at 9.8063. These results show and improvement of 500% in performance when using FPGA versus a PC due to the massive parallelism of the FPGA implementation. However, in order to utilize the full performance of the GA, it must be either controlled by a real time operating system, or at least by a C program running as a kernel module.

5 Conclusion and Future Work

In this work, we have designed, simulated and verified the functionality of a genetic algorithm to solve a 4-way NP complete partitioning problem of 128 16-bit unsigned integers. GA operators and parameters were selected based on Matlab simulations. The GA was implemented using VHDL and Altera's IP cores and verified using SystemVerilog with ModelSim. The design was then tested on Terasic's SoCKit as a coprocessor to accelerate the algorithm's execution on the embedded ARM Cortex-A9 MPCore processor. Test results demonstrate that using the massive parallelism of FPGAs, it is possible to achieve multiple times higher performance while using a fraction of the size and energy consumption of modern desktop computers. We are currently working on the comprehensive functional verification of the proposed design. In the future, we are planning to use a combination of standard functional coverage and rigorous formal verification techniques to meet industrial verification standards. GA could also be used in software testing (Mantere and Alander, 2005) and the objective could be e.g. the verification of the recently proposed flexible floating point numbers called unums (Gustafson, 2015).

References

Jarmo T. Alander. Indexed bibliography of genetic algorithms genetic algorithms and evolvable hardware and FPGAs. *Report 94-1-FPGA*. University of Vaasa, Dept of Information Tech and Production Econ, 2008. http://www.uwasa.fi/ TAU/reports/report94-1/gaFPGAbib.pdf

Jarmo T. Alander, Mikael Nordman, and Henri Setälä. Register-level hardware design and simulation of a genetic algorithm using VHDL. *In Proceedings of the MENDEL'95*, 10-14, 1995.

Jarmo T, Alander. On optimal population size of genetic algorithms. *In Proceedings of Comp Euro 92, Computer Systems and Software Engineering*, 65-70, 1992. doi:10.1109/CMPEUR.1992.218411.

Peter Alfke. *Efficient Shift Registers, LFSR Counters, and Long Pseudo Random Sequence Generators*. Xilinx Corporation, 1996.

Adriel Cheng and Cheng-Chew Lim. Optimizing system-on-chip verifications with multi-objective genetic evolutionary algorithms. *Journal of Industrial and Management Optimization*, 10(2):383-396, 1996. doi:10.3934/jimo.2014.10.383

Shi-Yi Gao, Xiao-Hua Luo and Yu-Feng Lu. Functional coverage convergence technique based on genetic algorithm. *Journal of Zhejiang University (Engineering Science)*, 49(8):1509-1515, 2015. doi:10.3785/j.issn.1008-973X.2015.08.015

John L. Gustafson. *The End of Error: Unum Computing*, 2015. CRC.

Brian Hayes. The Easiest Hard Problem. *American Scientist*, 90(2):113-117: 2002. doi:10.1511/2002.2.113.

Richard E. Korf. Multi-Way Number Partitioning. *In Proceedings of IJCAI'09 the 21st International Joint Conference on Artifical Intelligence*. 538-543, 2009.

Timo Mantere and Jarmo T. Alander. Evolutionary software engineering, a review. *Applied Soft Computing*, 5(3):325-331, 2005. doi:10.1016/j.asoc.2004.08.004

M. Saqib Nawaz , M. IkramUllah Lali and M. A. Pasha. Formal verification of crossover operator in Genetic Algorithms using Prototype Verification System (PVS), *In Proceedings of the 2013 IEEE International Conference on Emerging Technologies (ICET)*, 2013. doi:10.1109/ICET.2013.6743532

S. Owre, J. M. Rushby and N. Shankar. PVS: A Prototype Verification System. *11th International Conference on Automated Deduction (CADE)*, Lecture Notes in Artificial Intelligence, Springer-Verlag, 1992. ISBN 978-3-540-55602-2.

Terasic. *SoCKit User Manual*, 2015.

Bruce Wile, John Goss and Wolfgang Roesner. *Comprehensive Functional Verification: The Complete Industry Cycle (Systems on Silicon)*. Morgan Kaufmann Publishers Inc, 2005. San Francisco, CA, USA.

Xillybus. *Xillybus IP cores and design services*, 2017. <http://xillybus.com/>.

Xillybus. *Licensing*, 2017. <http://xillybus.com/licensing>

Investigation of Robotic Material Loading Strategies using an Earthmoving Simulator

Eric Halbach Aarne Halme Ville Kyrki

Department of Electrical Engineering and Automation, Aalto University, Finland
eric.halbach@gmail.com,{firstname.lastname}@aalto.fi

Abstract

A kinematic earthmoving simulation environment was used to investigate job planning strategies which could increase the performance of automated material loading with a robotic compact skid-steered wheel loader. One new problem studied was the subdivision of a larger rectangular workspace using the smaller rectangular Scoop Area (SA). Two methods for selecting scooping approach vectors were also compared: a Zero Contour (ZC) method which assesses all possible perpendicular approaches along the bottom of the slope, and the proposed alternative High Point (HP) method which scoops towards the highest point in the current workspace from a fixed point. Three jobs were simulated to determine which scooping method and SA dimensions resulted in the highest excavation rate in a truck loading scenario. Assuming the same scoop filling effectiveness, the HP method was found to offer a higher rate than the ZC method due to its more limited driving envelope. The maximum HP rates were achieved with SA dimensions which were narrower and longer than with the ZC method, while the optimal SA dimensions were also found to be dependent on the job parameters. When a higher amount of material to excavate per area was present, smaller SAs resulted in higher rates.

Keywords: automation, robotics, earthmoving, excavation, wheel loader, simulation, job planning

1 Introduction

Using robotic earthmoving machinery at mining and construction sites offers the possibility of both increasing safety and lowering costs. By separating human operators from the worksite, exposure to potential hazards such as collisions, rockfalls, dust and fumes is reduced, while commuting times can also be cut by controlling machinery from the safety and comfort of an office, which could be located far from the site.

This leads to the question of how such robotic machines would be controlled. By automating parts of the load-haul-dump work cycle and limiting direct teleoperation, efficiency can be increased by allowing one human to monitor and/or control several machines. Some commercial systems such as Sandvik's AutoMine and Caterpillar's Command for Underground already make this a reality by automating the hauling and dumping segments of the work cycle, though the loading or excavation phase usually needs to be controlled by a skilled human operator, possibly remotely by teleoperation.

Automating the excavation or loading phase could further increase efficiency by enabling fully automated work cycles. This is made difficult, however, by the unpredictable reaction forces encountered in ground material, which can contain fragments of unknown sizes and behave differently depending on factors such as compaction and moisture. Despite this challenge, solutions have been proposed for autonomous bucket filling, with some demonstrations being performed using full-sized machinery (Lever and Wang, 1995; Sarata et al., 2008; Almqvist, 2009; Dobson et al., 2015).

Fully automated work cycles could also make systems applicable in situations where direct teleoperation is not possible due to a long telecommunication time delay, such as in some space applications (e.g. 4-21 minutes to Mars one-way). Earthmoving capabilities on other planetary bodies would be useful for establishing a permanent human presence, for jobs such as settlement construction and harvesting regolith for resource extraction (Mackenzie et al.; Petrov, 2004). Given the additional hazards of radiation exposure and risk of depressurization which humans would face operating in these environments, supervising fully automated robots from Earth may be the ideal case for such jobs. Even if humans are located on site, however, full automation would be desirable for reducing human workload and freeing the crew for other important tasks.

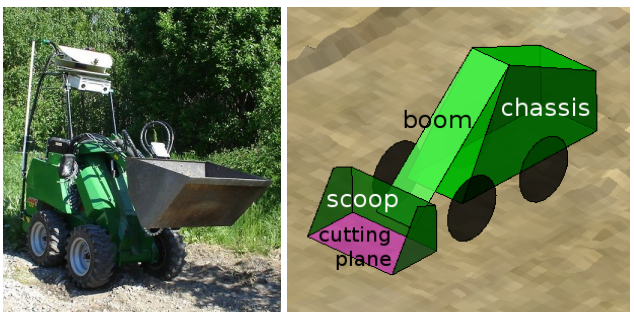


Figure 1. Compact skid-steered Avant 320 (*left*) and virtual model (*right*).

Assuming that scooping can be controlled automatically, a higher-level planning requirement for automated

earthmoving is deciding *where to dig* within a designated workspace such that progress is made towards the goal state. It may also be desirable to optimize some criterion, such as maximizing the excavation rate or minimizing energy use. This paper presents simulations for investigating this problem in the case of automated material loading by a robotic compact skid-steered wheel loader, modeled after an Avant 320 which was available to the authors for testing (see Figure 1).

The next section begins by presenting related work in this research area, followed by a description in Section 3 of the simulation environment used. Section 4 presents the strategy developed for subdividing a large rectangular workspace using the smaller rectangular Scoop Area (SA) in a truck loading scenario. Two methods for generating scooping approaches are then described in Section 5: the Zero Contour (ZC) method which selects a perpendicular approach along the bottom edge of the slope, and the proposed alternative High Point (HP) method which scoops towards the highest point in the workspace from a fixed point. Section 6 presents simulation results of jobs which were repeated using various SA dimensions, and both scooping methods, to observe the effect on the excavation rate. Different job parameters were used to also investigate the effect of less surrounding slope collapse and a higher slope. The conclusion and areas for future work are then discussed in Section 7.

2 Related Work

The work in this paper is partly based on the multi-resolution planning for robotic earthmoving developed by Singh and Cannon, which first subdivides a larger workspace with a coarse planner, then select digging locations with a refined planner (Singh and Cannon, 1998). Their planning for a wheel loader assumes the presence of an independently positioned dump truck which is to be filled, with the scooping actions limited to a region near the truck. Scooping actions are made perpendicular to the *zero contour*, or bottom edge of the slope, to achieve even loading of the bucket. All scooping locations are assessed before selecting one based on maximizing contour convexity into the scoop (to ease loading), minimizing side load (for an even fill) and minimizing the distance to the truck (Singh and Cannon, 1998).

Sarata et al. demonstrated automated scooping and truck loading cycles with a full-size wheel loader (Sarata et al., 2008). With their method, scooping locations are also located at the zero contour, with the scooping action oriented so as to minimize the predicted side moment on the bucket, to reduce wear. For the next action, the point a certain distance to the right or left of the previous one is chosen which is feasible and minimizes the hauling distance (Sarata et al., 2005).

Magnusson and Almqvist extend the work of Singh and Cannon for wheel loaders by using a more complex bucket model, and by evaluating convexity and side load along the entire bucket fill trajectory (Magnusson and Almqvist, 2011). Magnusson et al. also developed a coarse-to-fine planner and show how it ensures the long-term availability of good scooping locations as a large pile is excavated (Magnusson et al., 2015).

The ZC method implemented in this paper is a simplified 2D version of that proposed by Singh and Cannon (Singh and Cannon, 1998). It serves as an example of a method that selects from a large number of feasible actions along the contour, and is compared with the simple proposed HP method.

The second main planning investigation in this paper has so far not been found in the literature, though it has been alluded to (Singh and Cannon, 1998). I.e. in the case of dump truck locations which are dependent on excavating a slope evenly, which sub-region dimensions should be used to optimize some desired criterion (such as the excavation rate)?

3 Earthmoving Simulator

The robotic earthmoving strategies were investigated using a simulation environment developed using Matlab, based on previous work by the authors (Halbach and Halme, 2013; Halbach, 2007). It is similar to that used by Sarata and Magnusson et al. (Sarata, 2001; Magnusson et al., 2015), and allows ground material to be removed and deposited while maintaining a maximum angle of repose and conserving the total volume of material (thereby assuming a constant material density).

The simulator is purely kinematic and does not model forces, an approach taken for simplicity and because it was not intended for developing control of scooping actions, but rather for developing high-level planning strategies such as *where to dig* and *where to dump* material as a worksite changes over time. It therefore offers a compromise between the physics-based approach used in other simulators (Bonchis et al., 2011; Schmidt et al., 2010; Pla-Castells et al., 2009), and simulators developed primarily for visualization of construction processes which do not necessarily conserve the amount of material (Kamat and Martinez, 2005; Lipman and Reed, 2000).

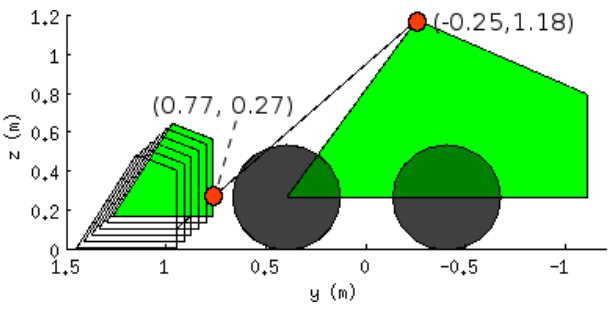


Figure 2. Avant model kinematics and range of scooping configurations resulting from extension of prismatic boom joint.

In this environment, a worksite is modeled as a surface with a 0.1 m grid resolution. The Avant 320 model

(see Figure 1) has wheels spaced 0.79 m width-wise and axles spaced 0.80 m apart, with the wheels, rendered as 2D disks, representing the centres of the tires. The vehicle's location and heading angle are defined in the XY-plane, while the current pose is determined from the average height of the four surface points at the 2D wheel locations, and the average slope between these points.

Machines in the simulator are assumed to possess accurate positioning and autonomous driving capabilities. Driving occurs by turning on the spot at a rate of 30 °/s and following straight paths at 0.5 m/s. One timestep in the simulator is 1/3 s.

The scoop has dimensions 0.89 m wide by 0.5 m long, and a volume capacity set at 0.15 m^3. Three joints are available for scoop positioning: rotary, between the chassis and boom; telescopic, for extending the boom; and rotary, between the boom and bucket. Figure 2 shows kinematic details of the joint locations, with the vehicle (reference point in the middle of the wheels at ground level) at $y = 0$, and the boom and bucket in their home positions.

Scoop-ground interaction works by checking for intersections between the bottom "cutting plane" of the scoop and the ground surface at each time step. Figure 3 (left) shows how the cutting plane is discretized with the red circular points. If any of the blue square ground points are above the corresponding point in the cutting plane (as in Figure 3, middle), the ground point is lowered and the column volume above added to the scoop load.

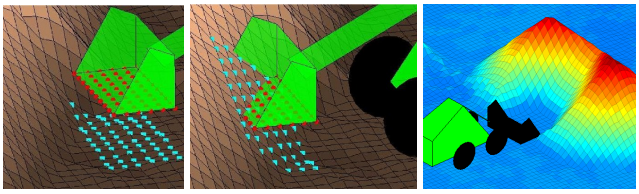


Figure 3. Checking for intersection of discretized cutting plane points (red circles) with corresponding ground points (blue squares) during scooping action (*left, centre*); result of simulated slope collapse (*right*).

Scooping actions are performed with the cutting plane level (both rotary joints in home position), with a boom extension ranging from 0 m to 0.24 m, corresponding to the bottom of the scoop positioned 0.17 m to 0.01 m above the ground (see Figure 2). The value to use for the next action is determined automatically at the end of each current action by comparing the scoop height with the desired ground level. If the scoop ended up too high or low, the boom setting is extended or retracted accordingly by 0.01 m for the next scooping action. Although the next action is usually at a different location, this strategy generally helps to maintain the designated scooping area at ground level. This strategy is necessary due to the kinematic nature of the simulator, which allows ground heights to be lowered by any intersection with the cutting plane. In a dynamic environment this strategy may not be necessary since the bucket could collide with and/or scrape along the ground, and a constant preset scooping configuration might be possible.

During a scooping action, material is added to the scoop load until the time step at which the current volume increment would cause the scoop capacity to be surpassed. At this point, a certain minimum fill ratio is assumed, with the remaining scoop capacity filled randomly and any leftover material deposited back on the ground. With a minimum of 0.8, for example, an average of 0.9 results over many actions. This strategy was developed so that it would be possible to specify the average performance of the scooping controller, which is assumed to exist, while allowing for some random effects due to tool-ground interaction.

When the bucket is raised after a scooping action, slope collapse is simulated by scanning in the X and Y directions for slope sections which are above the maximum specified repose angle. These are then adjusted (conserving volume), and neighbouring sections checked, until stability is reached. Figure 3 (right) shows the result of this simulated soil behaviour after several scooping actions into a pile. It is assumed that after any changes to the surface occur, the ground model would be updated by onboard laser scanners or by other surveyor robots.

4 Workspace Subdivision with Scoop Area

If excavating material from a large area with a wheel loader, the best coarse planning strategy may depend on where the material is to be deposited. If the dumping location is a stationary bin, then the loader would be free to select any location along the entire dig face - the scenario studied by Magnusson et al. (Magnusson et al., 2015). If dump trucks are being loaded, then a smaller digging region near the truck should generally be used to reduce the amount of driving between digging actions. Singh and Cannon studied the case of an independently positioned truck (Singh and Cannon, 1998), however here it is assumed that the main requirement is to excavate the slope face evenly, with the dump trucks positioned as needed to accomplish this goal.

Another assumption is that the workspace is rectangular, thus the method followed to excavate the slope evenly is to scan the workspace in a raster pattern from front to back with the smaller rectangular Scoop Area (SA), shown in Figure 4 mid-way through Job 1a. This job consists of excavating an 11 x 2.4 m section out of a 0.87 m-high plateau with a 30° slope. When a location is found which has ground heights a certain threshold (here 0.15 m) above ground level, the loader works there until the SA is cleared, and the SA then scans for the next location, with the machines repositioning there.

The graphical objects rendered in Figure 4 are interactive planning tools developed previously by the authors (Halbach and Halme, 2013). The large rectangular surface is used to specify and visualize the full workspace, while the triangular prism marks the Approach Side (AS).

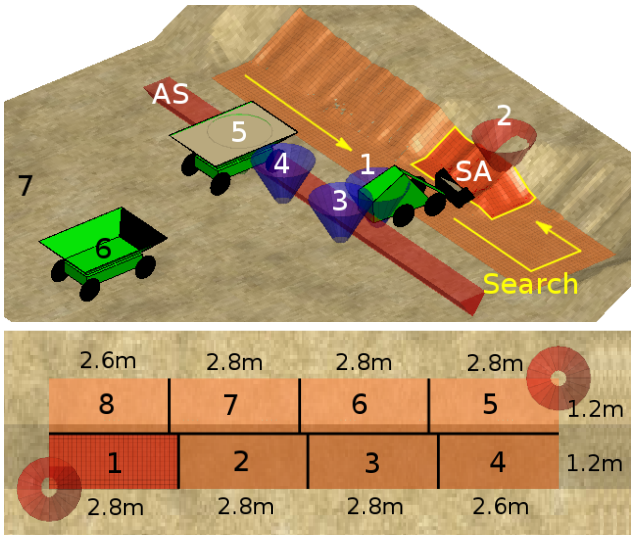


Figure 4. Scoop Area (SA) scans workspace for next working location along raster pattern from front to back; SA target dimensions (here 3 x 1 m) adjusted to divide into workspace (11 x 2.4 m) evenly.

A scooping action begins at the Stage point (cone 1) and is directed towards the Scooping Destination (cone 2), with the loader reversing to point 1 after the load is extracted.

Cones 3 and 4 represent driving waypoints for load transfer to the dump truck (at cone 5), which has a load capacity of 1 m^3 and is also skid-steered. A 2nd dump truck waits at point 6, and when one truck is filled, it drives to point 7 where the load is deleted, and continues to point 6, while the other truck drives to point 5. Points 6 and 7 would be the end of a hauling road along which the loads are transported, though this is not included here. The points are positioned relative to the current SA location. It should be noted that these planning strategies are specific to skid-steered machines which can turn on the spot.

The SA scans for the next working location with steps of one width and length. Its intended "target" dimensions are sometimes altered by an algorithm which attempts to divide the full workspace by the SA dimensions evenly, to avoid SA locations which only contain a fractional amount of work. The bottom of Figure 4 shows how the workspace is divided using target dimensions of 3 x 1 m.

5 High Point and Zero Contour Methods

This section describes the two methods for generating scooping approach vectors which are compared. The ZC method is based on the work of Singh and Cannon (Singh and Cannon, 1998), and selects a perpendicular approach along the bottom edge of a slope after evaluating convexity and the distance to point 3 (see Figure 4). A zero contour is first constructed by searching the workspace for points a certain threshold (here 0.15 m) above ground level, then following the contour until either the edge of the workspace is reached or the contour is closed. Figure 5 shows an example of contours constructed around an irregular pile shape.

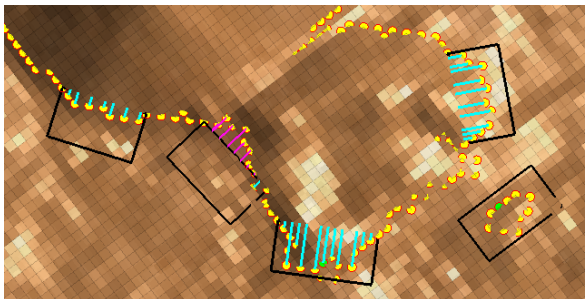


Figure 5. Convexity evaluation at possible scooping locations along zero contour; small separate contour at right assigned one possible location.

All possible scooping locations are then assessed by tracing along the contours with a line segment as wide as the scoop, with each end of the segment touching the contour. The convexity at each location is determined in 2D by adding the perpendicular line segments of points in between which protrude past the line, and subtracting those beyond the line (blue and purple lines in Figure 5). A possible scooping location is selected if its convexity is over 10% greater than the best value found so far (to attempt increasing filling effectiveness), or if it is within 10% of the best value and closer to point 3 (to reduce driving). Approaches which have a backwards-facing heading are not considered (to avoid excessive maneuvering), nor are those with non-traversable paths. If no acceptable scooping locations can be found, the HP method is used as a backup (described next).

In Figure 6 the ZC method is being used to excavate an SA location in Job 1a, with the yellow points showing the contour. At left is a new SA location, with a scooping approach selected at the corner which maximizes convexity. At right an unloading action is shown, and the different shape of the contour is also evident after 9 actions.

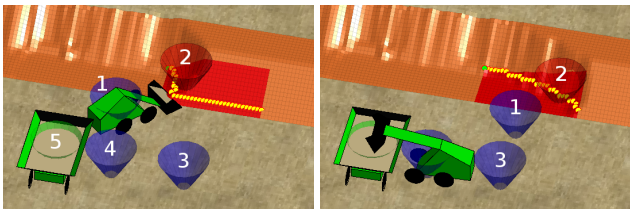


Figure 6. Excavation of SA with ZC method, with new Stage point (1) and Scooping Destination (2) for each scooping action; (*left*) first action at new SA, (*right*) after 9 actions with unloading at truck illustrated.

The HP method is a simple alternative which was proposed in previous work by the authors (Halbach and Halme, 2013), whereby the Stage point remains stationary and scooping actions are directed towards the highest point in the SA (see Figure 4). This results in a fan-shaped

pattern as the highest point shifts due to slope collapse, illustrated at left in Figure 7. In this example from Job 1a, 18 actions were needed to level the SA location. This coverage pattern can be compared with that using the ZC method at right, which consisted of 19 actions to clear the SA. The ZC driving paths appear to require more turning and driving, since they approach the SA more from the left side.

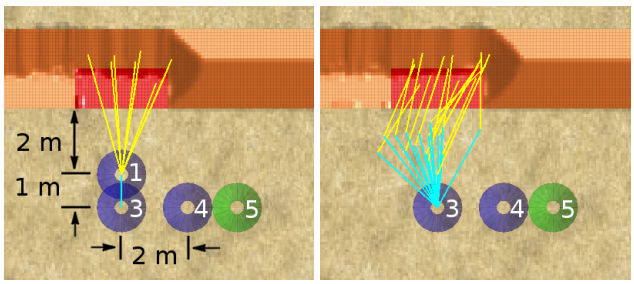


Figure 7. Coverage pattern for leveling 2.8 x 1.2 m SA location with HP method (*left*) using 18 actions, and ZC method (*right*) using 19 actions.

The HP method was not originally intended to be an improvement over others that have been proposed, but was meant to be a simple way of generating commands in order to test excavation jobs in the simulator. Its real-world effectiveness, which would need to be tested, may be hindered by the fact that it does not consider contour convexity or side loading, and can result in non-perpendicular approaches into the slope. If it could work well enough in practice, however, it may offer the benefit of reduced total driving, which is investigated next.

6 Simulation Results and Discussion

To find the SA dimensions which result in the maximum excavation rate, Job 1a was repeated with various SA widths and lengths, using both scooping methods. The minimum scoop filling ratio was kept constant at 0.8 (average fill of 0.9), thus a main assumption is that both the HP and ZC methods perform with the same scooping effectiveness.

A constant plateau height was chosen for this job in an attempt to reduce the factors which could affect the excavation rate, so that the SA dimensions would be the main variables during each simulation. With this constant-height plateau, each row excavated should have the same amount of material collapsing in from uphill, though the amount collapsing from the sides would initially increase as the front slope is excavated.

Two more versions of Job 1 were simulated to observe the effect of less surrounding slope collapse and a higher plateau (see Figure 8). Job 1b (at left) is a stand-alone plateau which fits in the same 11 x 2.4 m workspace, and has the same height and slope angle. Job 1c (at right) is similar to Job 1a, with the same workspace dimensions, however with the plateau height doubled to 1.73 m.

The excavation rate results for these jobs are plotted in

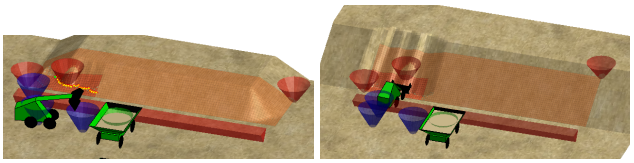


Figure 8. Job 1b (*left*), 0.87 m-high stand-alone plateau or pile, and Job 1c (*right*), 1.73 m-high plateau, both with 30° slopes.

Figure 9. Included in these plots are the ranges of SA dimensions which were chosen manually to show the rise and fall of the rate, with the width on the X-axis and different lines plotted for each length value. Each point is the average rate recorded after repeating the job 10 times, which was assumed to be sufficient given the randomness introduced in the scoop filling. The error bars represent one standard deviation.

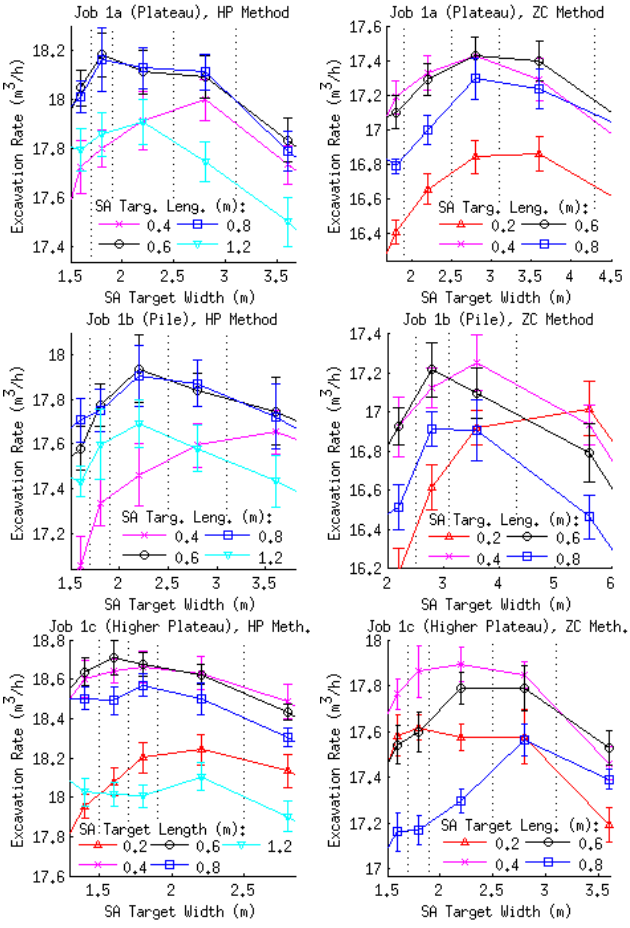


Figure 9. Excavation rate for Jobs 1a-c with varying SA target dimensions using HP and ZC method; 0.8 minimum scoop load ratio; 10 trials per data point.

In these plots, each data point represents a range of target SA dimensions which map to that value due to the algorithm which attempts to divide the workspace evenly. For the width dimension, the ranges are represented by the dotted lines. Target SA widths of 2.0-2.4 m, for example, map to a width of 2.2 m (5 SA locations along 11 m width). Similarly, a line for the target length of 1.0 m is

Table 1. Maximum Excavation Rate (Rate 1) and Volume per Combined Drive Time (Rate 2) for Jobs 1a-c

Job	HP Method			ZC Method		
	Max. Rate 1 (m^3/h)	Target Width (m)	Target Length (m)	Max. Rate 1 (m^3/h)	Target Width (m)	Target Length (m)
1a	18.182	1.8	0.6	17.433	2.8	0.4
1b	17.938	2.2	0.6	17.253	3.6	0.4
1c	18.710	1.6	0.6	17.891	2.2	0.4
	Max. Rate 2 (m^3/h)	Target Width (m)	Target Length (m)	Max. Rate 2 (m^3/h)	Target Width (m)	Target Length (m)
1a	15.519	2.8	0.8	15.003	3.6	0.6
1b	15.148	2.8	0.8	14.722	5.6	0.6
1c	15.999	2.2	0.6	15.458	2.8	0.6

not plotted since this is mapped to 1.2 m. The maximum rates and corresponding SA dimensions for each method and version of Job 1 are summarized in Table 1.

The table also includes results with the *volume per combined drive time* measure (plots not shown), which is the volume excavated divided by the total driving and turning time of the loader and two dump trucks. This is included to consider the case where excavating with minimal driving would be more important than excavating quickly, e.g. if energy is limited such as in a planetary construction scenario.

One observation is that in each case, the HP method achieves a higher maximum rate than the ZC method, likely due to the HP method's more limited coverage pattern with less turning and driving. This, again, assumes the same bucket filling effectiveness for both methods.

Another observation is that with the HP method, the maximum rates are achieved with SAs which are narrower and longer than with the ZC method. One reason for this could be that after unloading at the truck, with the HP method the loader always turns 90° at point 3 to reach point 1 (see Figure 7), therefore narrower SAs may be preferred to reduce further turning. The ZC method may prefer shorter SAs because they cannot contain contours with much curvature, and more curvature could result in more maneuvering to approach from the side. Short SAs would then need to be wider to contain enough material so that the SA does not reposition too frequently, which increases driving. Since with the ZC method the loader turns at point 3 by varying amounts towards the moving Stage point, far ends of wider SAs can perhaps be reached sooner than with the HP method.

It can also be observed when comparing the different job versions that when there is more material to excavate per area, such as with the higher plateau of Job 1c compared with Job 1a, or with more surrounding slope collapse in Job 1a compared with Job 1b, higher rates result. These are also usually achieved with smaller SAs, likely because with more material to excavate per area, smaller SAs become beneficial since they can be covered with less driving. Smaller SAs have the disadvantage of more repositioning of the machines between SA locations, however this is evidently outweighed by the advantage of less driving within the SAs.

Finally, the optimal SA dimensions with the volume per combined drive time measure are larger than with the standard excavation rate. This could be expected, since although larger SAs require more driving within them, they also require less repositioning between SAs. Repositioning would impose a bigger penalty with this measure since it involves all three machines driving simultaneously.

7 Conclusions and Future Work

The simulation results presented in this paper showed that the HP method resulted in higher excavation rates than the ZC method for various slope excavation jobs. One area for future work would be to check if a real loader could indeed fill its bucket as effectively with the HP method, despite the possible drawback of occasional non-perpendicular approach vectors which could result in asymmetrical loading. Future work would also include implementing the system with robotic machines and demonstrating the necessary site modeling, autonomous driving and scooping control capabilities.

The ZC method tended to reach its maximum rates with SA dimensions which were wider and shorter than with the HP method. It was also found that when more material was present per area, due to a higher plateau or more surrounding slope collapse, smaller SAs resulted in higher excavation rates. For reducing total machine driving, larger SAs were beneficial.

In general applications, piles and slopes could have irregular shapes and heights, thus as another area for future work, an algorithm could be developed which first analyzes the properties of the slope to excavate, then estimates optimal SA dimensions. During the job, the dimensions could be adjusted automatically based on the observed slope properties, or also in a speculative way to see if a higher rate can be achieved.

Acknowledgment

The authors would like to acknowledge the Academy of Finland for funding the Centre of Excellence in Generic Intelligent Machines (GIM) (2008-2013), of which this research was a part, and also the Graduate School in Electronics, Telecommunications and Automation (GETA) for supporting this work (2011-2013).

References

Håkan Almqvist. Automatic bucket fill. Master's thesis, Linköping University, 2009.

Adrian Bonchis, Nicholas Hillier, Julian Ryde, Elliot Duff, and Cédric Pradalier. Experiments in Autonomous Earth Moving.

In *18th International Federation of Automatic Control (IFAC) World Congress*, Milan, Italy, 2011.

Andrew A. Dobson, Joshua A. Marshall, and Johan Larsson. Admittance Control for Robotic Loading: Underground Field Trials with an LHD. In *Field and Service Robotics*, Toronto, Canada, 2015.

Eric Halbach. Development of a Simulator for Modeling Robotic Earth-Moving Tasks. Master's thesis, Helsinki University of Technology, 2007.

Eric Halbach and Aarne Halme. Job planning and supervisory control for automated earthmoving using 3D graphical tools. *Automation in Construction*, 32:145–160, 2013.

Vineet R. Kamat and Julio C. Martinez. Large-Scale Dynamic Terrain in Three-Dimensional Construction Process Visualizations. *Journal of Computing in Civil Engineering*, 19(2): 160–171, 2005.

Paul J. A. Lever and Fei-Yue Wang. Intelligent Excavator Control System for Lunar Mining System. *Journal of Aerospace Engineering*, 8(1):16–24, 1995.

Robert Lipman and Kent Reed. Using VRML in Construction Industry Applications. In *5th Symposium on Virtual Reality Modeling Language (VRML)*, Monterey, U.S.A., 2000.

Bruce Mackenzie, Bart Leahy, Georgi Petrov, and Gary Fisher. The Mars Homestead: a Mars Base Constructed from Local Materials. In *Space 2006*, San Jose, U.S.A.

Martin Magnusson and Håkan Almqvist. Consistent Pile-Shape Quantification for Autonomous Wheel Loaders. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, San Francisco, U.S.A., 2011.

Martin Magnusson, Tomasz Kucner, and Achim J. Lilienthal. Quantitative Evaluation of Coarse-to-Fine Loading Strategies for Material Rehandling. In *IEEE International Conference on Automation Science and Engineering (CASE)*, Gothenburg, Sweden, 2015.

Georgi I. Petrov. A Permanent Settlement on Mars: The First Cut in the Land of a New Frontier. Master's thesis, Massachusetts Institute of Technology, 2004.

Marta Pla-Castells, Ignacio García-Fernández, Miguel A. Gamón, and Rafael J. Martínez-Durá. Interactive earthmoving simulation in real-time. In *Congreso Español de Informática Gráfica (CEIG) (Spanish Congress of Computer Graphics)*, San Sebastián, Spain, 2009.

Shigeru Sarata. Model-based Task Planning for Loading Operation in Mining. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Maui, U.S.A., 2001.

Shigeru Sarata, Yossewee Weeramhaeng, and Takashi Tsubouchi. Planning of scooping position and approach path for loading operation by wheel loader. In *22nd International Symposium on Automation and Robotics in Construction (ISARC)*, Ferrara, Italy, 2005.

Shigeru Sarata, Noriho Koyachi, and Kazuhiro Sugawara. Field Test of Autonomous Loading Operations by Wheel Loader. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nice, France, 2008.

Daniel Schmidt, Martin Proetzsch, and Karsten Berns. Simulation and Control of an Autonomous Bucket Excavator for Landscaping Tasks. In *IEEE International Conference on Robotics and Automation (ICRA)*, Anchorage, U.S.A., 2010.

Sanjiv Singh and Howard Cannon. Multi-Resolution Planning for Earthmoving. In *IEEE International Conference on Robotics and Automation (ICRA)*, Leuven, Belgium, 1998.

Modeling and Simulation as Support for Development of Human Health Space Exploration Projects

Agostino G. Bruzzone Marina Massei

DIME, University of Genoa, Italy,{agostino,massei}@itim.unige.it

Giuseppina Mùrino Riccardo Di Matteo Matteo Agresta Giovanni Luca Maglione

Simulationteam, Italy,
{giuseppina.murino;riccardo.dimatteo;matteo.agresta;giovanniluca.maglione}
@simulationteam.com

Abstract

The capability of the space environment to alter the cells behavior seems to be an opportunity for future researches in biology, for diseases such as cancer. This paper highlights the importance of Interoperable Simulation Systems as precious instruments to support and improve space exploration projects devoted to biological researches. The research investigates the potential of Modeling & Simulation to reproduce a virtual environment to support Nano-satellite experiments in cooperation among the different stakeholders involved in a space mission, such as scientist, engineers and biologists.

Keywords: modeling & simulation, HLA, space exploration; human health; cancer development and progression

1 Introduction

The sphere of the human environment and exploration continues to expand towards space; there is a need to enrich the knowledge on the effect of the Sun and "space weather" to preserve the safety of the astronauts (Marhavilas, 2004). The space environment conditions are extremely challenging for human body; indeed microgravity condition are joined with ionizing radiation sources including cosmic rays and solar particle events (SPEs) (Benton et al., 2001; Townsend, 2005). Several studies conducted on astronauts after they spent several months in space proved that an extended exposure to microgravity conditions are correlated with health problem, for example bone loss (Lloyd et al., 2008; Lang et al., 2004)

Despite the health of the astronauts, the space weather is a critical issue, its capability to alter the cells behaviour seems to be an opportunity for future researches in biology, for diseases such as cancer. This is proved by several studies, present in literature, describing experiments on microgravity devoted to gain insights into its effect on living organisms. (NASA, 2001).

Furthermore, several promising experiments have found some correlation with the behavior of certain cells and bacteria and the simultaneous conditions of microgravity joined with ionizing environment. The experimental results suggest that cell development and proliferation is different in microgravity conditions and within ionizing environment (Massimiani et al., 2014; Leys et al., 2009; Vanhavere et al., 2008; Mastroleo et al., 2009).

Some scientist supposes that microgravity conditions may give the possibility of developing tissues and biological samples in three dimensions as it happens within human bodies: analyzing the experiments conducted in a cell cultures in a 1-g environment the proliferation is evolving only in two dimensions, this does not make them perfectly representative of what actually happens inside our body where the growth is tridimensional. The same cell types in orbital systems highlights this substantial difference confirming that space is a unique and incomparable environment for biological research.

The reasons of this different behavior have not yet been fully determined, but it is supposed to be correlated to a number of investigated factors:

- Interaction with terrestrial magnetic field: it could cause other effects in addition to those caused by microgravity (considering the nature of membranes which act as electrical capacity);
- Microgravity: this element generates different behaviors in biomedical samples between real and simulated microgravity; indeed in the ground simulators it is not possible to reach the microgravity levels common in low-Earth orbits (on the range of 109 - 105 g). By altering gravity, we are able to investigate partially these effects on biological systems related to the presence and reaction to this unique force. However simulating microgravity on Earth for more than several seconds is impossible with existing technology. So by using spaceflights, we are starting to understand that not only gravity, but also the physical changes that occur in microgravity conditions, may have effects on the evolution of species and their ecologies

- Joint effect of microgravity and ionizing environment: the impossibility to reproduce the effects of microgravity in a laboratory (due to technological limits) does not allow to consider its combined interactions with ionizing radiations. In addition, the space radiation is different from the one that we normally experience on Earth, such as x-rays or γ-rays. The combined effect could result in additional elements affecting the radiation hazard caused by exposure; these are usually acting by "changing of body systems functioning at all levels: from cellular up to organism". In facts the ionizing radiation exposure causes changing on body systems.

All these factors, combined with the reduction of costs in space missions, increase the interest of biomedical research in using the space as a laboratory for its studies. Therefore the use of simulation could guarantee a right first time approach in setting up the experimentation & testing.

2 Nanosatellites for Biomedical Experimentation

The world's first artificial satellite, the Sputnik 1, was put in space by Soviet Union in 1957. Since then, thousands of satellites have been launched into the space and are nowadays in orbit around the Earth (Lanius et al., 2013).

Last technological innovation, in use to support biomedical experimentations, is represented by CubeSat nanosatellite generation (Puig-Suari,et al., 2001). These miniaturized satellites for space research, usually launched by a carrier rocket or launch vehicle are made up of multiples of 10 x 10 x 11.35 cm cubic units (Bouwmeester and Guo, 2010). They have a mass of about 1 kilograms per unit and a Geocentric Orbit at Low Earth Orbit (LEO) altitude, usually <0.1 times Earth radius.

3 Modelling & Simulation Supporting Space Missions

Space missions are extremely costly and dangerous and Modeling & Simulation (M&S) is extremely useful to support engineering and reduce risks; indeed M&S allows to evaluate feasibility of experimentation in terms of equipment and technical solutions (Baxter, 2010)

Decision makers as well as project leaders usually face limited resources and rigid time constraints for space experimentation; so they need to test the feasibility of complex systems before realizing them in order to avoid unexpected problems. That's why M&S provides a strong support, particularly in the initial phase of a project giving to stakeholders a holistic view of the whole context (Montgomery, 2000)

In previous researches the authors performed several studies on M&S reproducing complex systems behavior in different fields including space (Bruzzone et al., 2016), logistics (Bruzzone et al., 2014), Intelligent Agent Computer Generated Forces (IA-CGF), disaster recovery in critical environment (Bruzzone et al., 2016), reproduction of intelligent behavior (Wooldrige and Jennings, 1995; Bruzzone et al., 2015), data & communications exchange among different entities (Bruzzone et al., 2013) and training (Bruzzone et al., 2011; Bruzzone et al., 2016). Despite all these areas are really different one each other, they have a common line because they reproduce complex problems where non-linear functions lead often to counter-intuitive behaviors on the system itself that evolves dynamically along the simulated timeline. Furthermore, the models allows to simulate conditions and situations that are often impossible to be reproduced in experiments on the Earth, both for the costs and for technological complexities.

To this end, the authors propose a simulation devoted to investigate all the operations required for setting the real experimentation of the Nano-satellite technology applied in space for studying the combined effect of microgravity and ionizing radiation on cancer cells affected by GBM (Glioblastoma Multiforme).

GBM is the most common form of malignant brain tumors with a median survival time of patients with less than one year. It represents 52% of all cases of primary brain tumor and 20% of all intracranial tumors. This type of cancer is actually treated by the best health facilities through the surgery and subsequent exposure to chemotherapeutic and radio therapies (Mahaley et al., 1989)

Because of its nature "multifaceted", complete surgical tumor excision is often very difficult and sometimes impossible without permanent damage in the patient. Because of the high incidence of this type of cancer and its characteristics it is necessary to deepen the knowledge of its pathogenesis by studying the behavior in various environments including the space. A better and detailed knowledge of cells behavior affected by GBM, should lead to identifying the causes that generate it or pharmacological remedies to counteract their evolution and development.

For this reason, the basic idea is referred to the biological effects of ionizing radiation and microgravity that could increase the chances of success of treatments and biomedical applications.

4 Nanosatellites for Space Experimentation

In this paragraph the overall architecture and nature of the CubeSat Systems are described. Indeed, it is

proposed a summary of systems, subsystems and components system that have to be considered for being simulated:

- Mechanical Systems and Structure: the satellite is made mainly of Aluminum 6061T6; it is usually considered, also in these nanosatellites, by using rapid prototyping design to adopt space qualified materials to build some structural components and to reduce its weight. Biological experiments should be hosted in special insulated containment module that are designed to be equipped with monitoring system.
- Observation System: an onboard observation system is required to monitor cell samples at various stages of the experiment.
- Thermal and Heat Exchange Systems: to keep alive the cells during all the mission phases in necessary to maintain a controlled environment (around 37 °C) despite the space extreme variation in terms of temperature. The thermal control system consists of sensors, insulating material, heat exchangers.
- Ionization and Radiations Phenomena: to obtain accurate scientific results, experimenters must understand the radiation environment during the duration of the experiment. A micro-dosimeter is necessary to measure the amount of ionizing radiation that is absorbed by the samples. The number and type of micro-dosimeters will depend on the required accuracy.
- Microgravity Environment: In order to understand microgravity conditions during the experiment, researchers need to use different kinds of accelerometers, with accuracy and dimensions compatible with mission goals. In particular two kinds of accelerometers are used. One for high level of acceleration during the launch, the other to measure microgravity conditions.
- Power Systems: an autonomous power system is needed to operate the whole systems including life support for keeping the cells alive during all mission phases. The nanosatellite uses solar cells and lithium batteries to provide power. The power is managed by a batteries charge regulator. During the integration and launch phases the nanosatellite is turned off and for this reason an umbilical connection with the launch site will provide power to the payload before the satellite release in orbit. Different solutions for umbilical connection need to be analyzed taking into account also the possible launch vehicle available.
- Telemetry, Tracking and Communications System (TT&CS): the TT&C act as the unique communication system once the SC is released from launch vehicle system. The TT&C communication uplink allow the mission control center ground station to upload command sequences in order to program all the SC operations using a UHF/VHF radio link and also for the monitoring of the platform, Telemetry.
- On Board Computer (OBC): the On Board Computer act as the brain of the system coordinating the function of all the subsystems. The main features of the OBC include the presence of two independent, but cooperative cores: one low power consumption microcontroller for the general management of the satellite (payloads, TT&C, specimen status, retrieving of data from sensors); the second core is an FPGA (Field Programmable Gate Array) for implementing specific tasks or generic systems also with IP cores (Intellectual Property) of third parties.
- The OBC is provided with several sensors on board and a 9 degree IMU (Inertial Measurement Unit) made by 3 axes magnetometer, 3 axes accelerometer and 3 axes gyroscope.
- Attitude and Orbit Determination and Control System (AODCS): this module supports the in-orbit control of the satellite by using on-board sensors and actuators. This module support the orbit corrections, attitude and spin rate control with high accuracy to ensure the communications link and to properly analyze the results of environment monitoring system (e.g. radiation level). The AODCS controls the positioning of the satellite in the selected orbit, in order to understand the results from the micro-dosimeter according to the known radiation exposure levels.
- On Board Software for Power Systems (OBSW): this system is compatible with OBC hardware and is based on a firmware that could be updated in terms of functions along the mission if necessary.

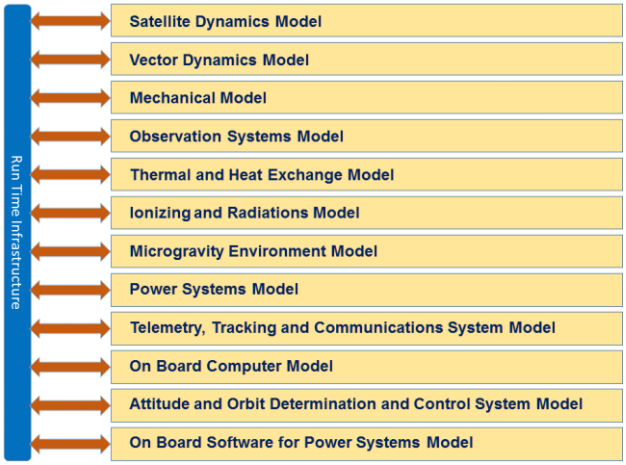


Figure 1. General Architecture.

5 General Architecture and Model Description

M&S supports strongly experimentation of space mission, giving the possibility to model the systems, subsystems and components and analyze interactions among them and biological samples.

The model is used to provide a measure of the resilience of the system to a hostile environment such as space, where reliability is challenging and it is necessary to evaluate redundancies, systems availability and capabilities.

Considering the flexibility demonstrated by the Intelligent Agents and Virtual Simulators developed within the MS2G (Modeling and Interoperable Simulation and Serious Games) paradigm (Bruzzone et al., 2014), the authors decide to realize a federation to be applied to Space Experimentation based on this approach. The authors benefits of their experience on past different simulators (Elfrey et al. 2011; Bruzzone et al., 2014; Bruzzone et al., 2016) created for SEE/Smackdown initiatives including SPIRALS (Space Interoperable Refilling and Advanced Logistics Simulator), IPHITOS (Interoperable simulation of a Protection solution based on light Interceptor Tackler operating in Outer Space) and SISMA (Medical Simulator of Astronaut including treatments, analysis and sickness models).

The simulator proposed in this case is designed to adopt the HLA IEEE1516 (High Level Architecture) standards and guarantees interoperability; the Federates could be based on stochastic models adopting combined simulation (continuous and discrete events combined together). The VV&A (Verification, Validation and Accreditation) for this simulation will be conducted along the entire FEDEP (Federation Development Process) by using SME (Subject Matter Experts) from Simulation Team (Bruzzone et al., 2014)

6 Description of the Different Models

In this paragraph all the models are described. It is important to outline that for the purpose of this simulator the biological specimen encapsulated in the Nano-satellite are considered as a "black box" representing a reference for the onboard systems in terms of temperature to be maintained and data to be collected. The simulation models include:

- Satellite Dynamics Model: the model regulates the physics of the satellite including motion and acceleration based on all its characteristics.
- Vector Dynamics Model: a specific model for the vector is included to reproduce the release process.
- Mechanical Model: Mechanical Systems devoted to release the CubeSat from the vector, the umbilical connections for power support during launch and interactions with movable parts.
- Observation Systems Model: simulates the sensors that are interacting directly with the black box constituting the biological specimen as well as the links of the sensors to the CubeSat Core Systems devoted to conduct measurements during the experiments on the specimen. This model should include failures and performance estimations related to the experimentation on the cells based on the data collection, boundary conditions, status of components and sensors.
- Thermal and Heat Exchange Model: considers the thermal effects and heat exchange in the CubeSat with special attention to the cells for the experimentation. The aim of this module is to control the temperature of the cells at a constant value of 37°C balancing the radiated heat from and to the CubeSat.
- Ionizing and Radiations model: is devoted to evaluate performances of the sensors respect the "solar weather" and the estimated exposure to the radiation of the specimen. In addition these models could reproduce the effect of radiations in terms of noise over the signals.
- Microgravity Environment model: it is the model of the sensors adopted to measure microgravity acceleration.
- Power Systems model (PS): the model deals with the computation of power absorbed and consumed/provided by the battery. It considers the power request to keep the temperature of the cell constant as well as consumptions due to sensor operations and communications. It considers then dynamic charge-discharge curve of the lithium battery in the cell depending from solar power and conditioning system consumption; this model coupled with the satellite dynamic model allows to estimate the exposure of the solar panels to the sunshine and their efficiency in the different asset configuration respect Sun and Earth relative positions.
- Telemetry, Tracking and Communications System (TT&CS) model: reproduces the datalink, a crucial component for correct operational profile. The cyber space is modeled to analyze the performances of the communication systems: it allows to visualize the information packet flow and evaluate communication system performances changing parameters characterizing the communication nodes and links in terms of availability, reliability and confidentiality. TT&CS are coupled with the dynamic evolution of the CubeSat around the Earth and respect the Ground Base that receive the experimental data to identify when/how communicate; the coupling with the Power System Model allows to consider the relative power consumption.

- On Board Computer model is crucial to implement and test on board computer system in a simulated environment before executing the mission. It also enables the possibility to evaluate the performance of Artificial Intelligence module (AI) devoted to direct dynamically the operation based on the level of decisional autonomy of the system during mission stochastic events.
- Attitude and Orbit Determination and Control System model (AODCS) simulates the performances of the sensors and modules devoted to control the orbit during the dynamic evolution of the mission. It could be used by reverse engineering to define the requirements of this system to achieve a specific overall performance.
- On Board Software for Power Systems model (OBSW) simulates the control firmware addicted to the control system for the Power System Module

Table 1. Description of the Simulation Parameters

Solar Exposition	**Dynamic Model**
Earth Sun distance [m]	Ground Station lat., long. & altitude
Earth diameter [m]	CubeSat latitude, longitude & altitude
Exposure Average angle [rad]	CubeSat Asset: Pitch, Yaw. Roll[rad]
Starting Shadow Angle [rad]	CubeSat linear and angular speeds
Ending Shadow Angle [rad]	CubeSat Status of Operations
Power coefficient [W/cm^2]	Vector latitude, longitude & altitude
Solar Panel Surface [cm^2]	Vector Asset: Pitch, Yaw. Roll[rad]
number of panels	Vector linear and angular speeds
Potential Solar Power [W]	
Current Efficiency of Solar Panels	**Thermal Model and Heat Exchange**
Current Battery Charging [mAh]	Solar Constant [W m^-2]
	Stefan's coefficient [W m^-2 K^-4]
	T Sunshine Side of the CubeSat [K]
SATCOM	T Shadow Side of the CubeSat [K]
max distance [m]	Emissivity Factor
max power consumption [W]	Radiation Factor
Gain transmitter [dBi]	CubeSat Surface [m^2]
Gain receiver [dBi]	Heat to be dispersed [W]
frequency [Hz]	Insulation factor
N Exponent for Env. Conditions	Absorbed Energy on Exposed Face [Wh]
Power receiver [W]	Dispersed Energy on shadow faces [Wh]
Communication Status	Current Energy Balance [Wh]
Bandwidth [Mb/s]	Battery Use for Heat Exchange [mAh]
Efficiency Level	
Data to be Transmitted [Mb]	**Battery**
Transmission Time [s]	Nominal Capacity [mAh]
Power Consumption [W]	Operational Voltage [V]
Energy Consumption [Wh]	Nominal Battery Energy [Wh]
Battery Use for SATCOM [mAh]	Initial State of Charge
	Current capacity Level [mAh]
Observation System Model	Current State of Charge
Dbase Capacity [Mb]	
Data to be Transmitted [Mb]	**On Board Computer Model**
Current Data Flow [Mb/s]	Availabilities of the different Systems
Observation System Status	Current Battery Use [mAh]

The models reproduce CubeSat dynamics in terms of orbit, power generation and consumption, battery re-charging and communication management. It considers heating exchange and temperature control of the nanosatellite from the launching moment.

Major model parameters are summarized in Table 1. The simulation aims to evaluate the performance with special attention to the power required to keep the internal satellite temperature at 37°C and communicate with the ground station. The total power is obtained by the interaction of different systems models (e.g. AODCS, OBC, TT&CS and PS). The energy to the different CubeSat systems is provided by battery and solar panels.

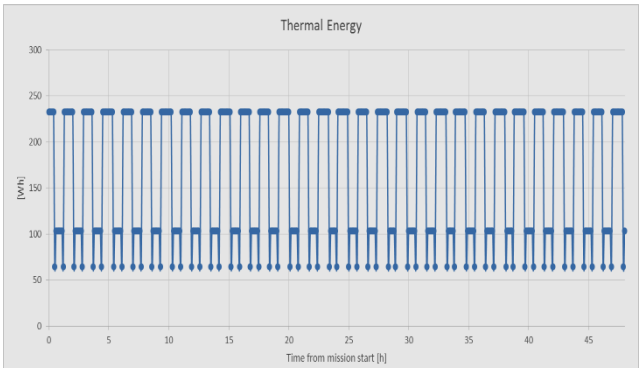


Figure 2. Generated Energy by Solar Panels

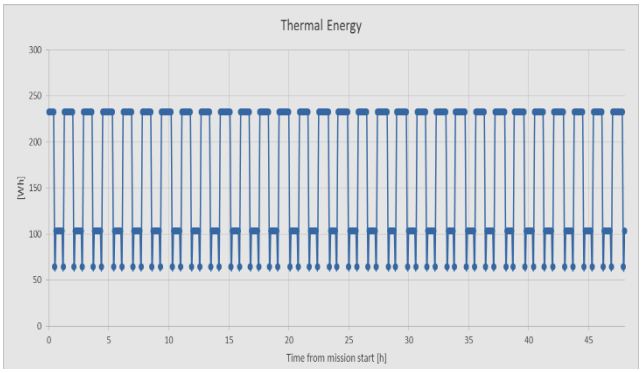


Figure 3. Thermal Energy Balance

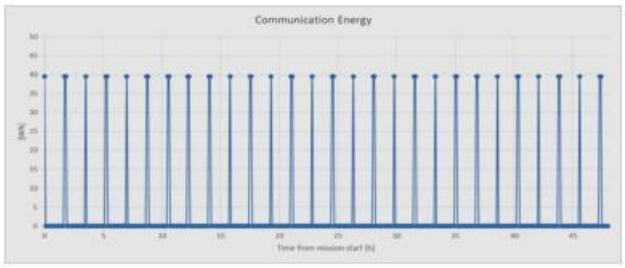


Figure 4. Energy required for Communications

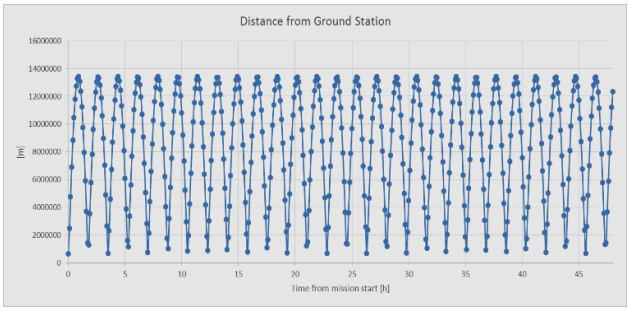


Figure 5. Communication Opportunities: CubeSat–Ground Base

The CubeSat is released by its vector when it is reached the proper operational orbit altitude; this process is activated by a simple spring mechanical device that releases it at a relative speed of 2.4 m/s perpendicular to the vector trajectory.

Obviously the model computes the CubeSat orbit respect the movements of Earth and Sun, in order to evaluate dynamically the solar exposure and the conditions of low efficiency when the solar panels are affected by atmosphere (CubeSat relative sunset and sunrise); this allows to estimate the power generated by the solar panels in different conditions. In general there are different situations including darkness, (no charge), shadow (partial charge) and full exposure (full charge).

Under the hypothesis of omnidirectional antenna, the cyber layer is modelled to analyze the conditions when communications are possible as well as the exchange of information with the ground station considering the relative orbital position of the satellite.

Indeed, the TT&CS model simulate all the data exchanges with the ground base, considering the different delay and the noise of the signal due to the position of the satellite. Since the CubeSat have a different rotation speed compared to the Earth, there are conditions where the satellite could not be able to transfer the data.

In these cases the TT&CS close the communication with the Ground Base in order to save power and the OBC stores the experimental data in its Dbase for late transmission opportunities.

7 Conclusions

In the paper is described a simulation model of the systems devoted to conduct an experiment on tumor cells under micro-gravity and radiation conditions on a nanosatellite in space.

The conceptual model has been developed and implemented in order to support design, engineering, virtual prototyping and risk analysis about the different systems.

The simulation resulted very useful to create a virtual prototype to deal with the complexity of the different systems and their interactions as well as with the large number of the variables; in addition, the stochasticity related to the potential failures affects the mission success and requires proper risk analysis to identify most convenient satellite configurations and redundancies to mitigate problems.

By this approach it is possible to reduce risks and guarantee success in this context dealing with nanosatellite, so a pretty compact system of systems that needs to operate in space guaranteeing success despite limited budget. Indeed the use of simulation allows to improve design and reduce overall costs and risk; obviously it is fundamental to proper model the different systems and their interactions to proper reproduce the different operations and conditions even considering potential failures and interferences.

This paper proposes the preliminary models developed and their general structure; currently the authors are working on the interoperability of this simulator to construct a federation able to guarantee that the final result of this research will be flexible and open to be integrated with other interoperable simulators. Indeed by this approach the simulator will be modular and available to be adapted for reproducing other similar space experiments.

References

M. J. Baxter. *Guidance for Space Program Modeling and Simulation.* June 30, AEROSPACE REPORT NO. TOR-2010(8591)-17, 2010.

E. R. Benton and E. V. Benton. Space radiation dosimetry in low-Earth orbit and beyond. *Nuclear Instruments and Methods in Physics Research*, section B, 184: 255-294, 2001.

J. Bouwmeester and J. Guo. Survey of worldwide pico-and nanosatellite missions, distributions and subsystem technology. *Acta Astronautica*, 67: 854-862, 2010.

A. Bruzzone. Intelligent Agents for Computer Generated Forces. Invited Speech at Gesi User Workshop, Wien, Italy, October 16-17, 2008.

A. Bruzzone, A. Tremori, and M. Massei. Adding Smart to the Mix. *Modeling, Simulation & Training: the International Defence Training Journal* 3: 25-27, 2011.

A. Bruzzone, D. Merani, M. Massei, A. Tremori, C. Bartolucci C., and A. Ferrando. Modeling Cyber Warfare in Heterogeneous Networks for Protection of Infrastructures and Operations. In *Proceedings of I3M'13 - EMSS*, Athens, September 25 – 27, 2013.

A. Bruzzone, F. Longo, M. Agresta, R. Di Matteo, and G.L. Maglione. Autonomous Systems for Operations in Critical Environments. In *Proceedings of the Modeling and Simulation of Complexity in Intelligent, Adaptive and Autonomous Systems (MSCIAAS 2016) and Space Simulation for Planetary Space Exploration (SPACE 2016)* Pasadena, California, April 03 - 06, 2016.

A. Bruzzone, F. Longo, M. Massei, L. Nicoletti, and M. Agresta. Safety and security in fresh good supply chain. *International Journal of Food Engineering*, 10: 545-556, 2014.

A. Bruzzone, L. Dato, and A. Ferrando. Simulation Exploration Experience: Providing Effective Surveillance and Defense for a Moon Base against Threats from Outer Space. In *Proceedings of IEEE/ACM 18th International Symposium on Distributed Simulation and Real Time Applications* 01 - 03 October 2014, France, Toulouse.

A. Bruzzone, M. Massei , M. Agresta, A. Tremori, F. Longo, G. Murino, F. De Felice, and A. Petrillo. Human behavior simulation for smart decision making in emergency prevention and mitigation within urban and industrial environments. In *Proc. of 27th EMSS European Modeling & Simulation Simposium*, 2015.

A. Bruzzone, M. Massei, A. Tremori, F.Longo, L. Nicoletti, S. Poggi, C. Bartolucci, E. Picco, and G. Poggio. MS2G: simulation as a service for data mining and crowd sourcing invulnerability Reduction. In *Proceedings of WAMS2014*, Istanbul, Turkey, September 16-19, 2014.

P. R. Elfrey, G. Zacharewicz, and M. Ni. Smackdown. Adventures in Simulation Standards. In S. Jain, R.R. Creasey, J. Himmelspach, K.P.White, and M. Fu (Eds.), *Proceedings of the 2011 Winter Simulation Conference*, 2011.

T. Lang, A. LeBlanc, H. Evans, Y. Lu, H. Genant, and A. Yu. Cortical and trabecular bone mineral loss from the spine and hip in long-duration spaceflight. *Journal of Bone and Mineral Research*, 19:1006-1012, 2004.

R. D. Lanius, J. M. Logsdon, and R. W. Smith (Eds.), *Reconsidering Sputnik: Forty years since the Soviet satellite*. Routledge. 2013.

N. Leys, S. Baatout, C. Rosier, A. Dams, C. s'Heeren, R. Wattiez, and M. Mergeay. The response of Cupriavidus metallidurans CH34 to spaceflight in the international space station. *Antonie van Leeuwenhoek International Journal of General and Molecular Microbiology*, 96: 227-245, 2009. DOI: 10.1007/s10482-009-9360-5. PMID: 19572210.

S. A. Lloyd, E. R. Bandstra, N. D. Travis, G. A. Nelson, J. D. Bourland, M. J Pecaut, and T. A. Bateman. Spaceflight-relevant types of ionizing radiation and cortical bone: Potential LET effect?. *Advances in Space Research*, 42: 1889-1897, 2008.

M. S. Mahaley Jr, C. Mettlin, N. Natarajan, E. R. Laws Jr, and B.B. Peace. National survey of patterns of care for brain-tumor patients. *Journal of neurosurgery*, 71: 826-836, 1989.

P. K. Marhavilas. *The Space Environment and its impact on human activity*. 2004.

C. Massimiani, S. Gemini Piperni, M. Carnio, W. Zambuzzi, C. Cappelletti, and F. Graziani. Space systems design for research on the interaction of osteoblast-like cells and biomaterials (hydroxyapatite particles and titanium) in microgravity environment. IAC-14-A1-3-6, In *Proceedings of the 64th International Astronautical Conference* Toronto, September 2014.

F. Mastroleo, R. Van Houdt, B. Leroy, M. Benotmane, A. Janssen, M. Mergeay, F. Vanhavere, L. Hendrickx, R. Wattiez, and N. Leys. Experimental design and environmental parameters affect Rhodospirillum rubrum S1H response to space flight. *International Society for Microbial Ecology*, 3: 1402-1419, 2009. DOI: 10.1038/ismej.2009.74.

D. C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, New York (NY) 2000.

J. Puig-Suari, C. Turner, and W. Ahlgren. Development of the standard CubeSat deployer and a CubeSat class PicoSatellite. In *Proc. of IEEE Aerospace Conference*, Vol. 1, pp. 1-347, 2001.

J. M. Quero, L. León, J. Jiménez, M. Brey, J.M. Moreno, S. de la Rosa, A. Sánchez, D. López, and C. Leyva. CEPHEUS, a Multi-project Satellite for technology qualification. *Acta Astronautica*, 117: 238-242, 2015.

L. W. Townsend. Implications of the space radiation environment for human exploration in deep space. *Radiation Protection Dosimetry*, 115: 44-50, 2005.

F. Vanhavere, J. L. Genicot, D. O'Sullivan, D. Zhou, F. Spurny, I. Jadrnickova, G. Sawakuchi, and E. G. Yukihara. DOsimetry of BIological EXperiments in SPace (DOBIES) with luminescence (OSL and TL) and track etch detectors. *Radiation Measurements*, 43: 694-697, 2008. DOI: 10.1016/j.radmeas.2007.12.002.

M. Wooldrige and N. R. Jennings. *Intelligent Agents: Theory and Practice*. 1995.

The Drawing of a New Era of Research, Research Results Accomplishments: An Analysis of Results from 2000-2001, NASA technical records Practice.

SDNizing the Wireless LAN - A Practical Approach

Manzoor A. Khan Patrick Engelhard Tobias Dörsch

DAI Labor, TU Berlin, Berlin Germany,
{manzoor-ahmed.khan,patrick.engelhard,tobias.doersch}@dai-labor.de

Abstract

The emerging Internet of Things (IoT) paradigm and a plethora of diverse applications provision more flexible network management. Software Defined Networking (SDN) occupies the pivotal role in realizing such flexible network management. However, the gain of this potential panacea is still unmeasurable in a real sense especially when wireless medium is part of the equation, as the validation frameworks mostly skip capturing realistic system dynamics. In this paper, we study the performance gain of SDN control implemented in a physical testbed comprising of a virtualized core and a WLAN access network. With this contribution, we aim at realizing a more realistic environment where the impact of system dynamics on the stakeholders (users and operators) may be studied. We developed a mechanism to map the logical wireless channels over the physical wireless interface of the access point. SDN (OpenDaylight) control application for mobility management, the mapper tool, a visualization and control GUI, and Android applications are amongst the main contribution of this work.

Keywords: software defined networking, OpenFlow, Open vSwitch, experiment automation

1 Introduction

The recent past has advocated a rapid evolution in mobile communication technologies, which enables the transition from a monolithic architecture to a shared architecture and ensures ubiquitous connectivity. Such transformation, when coupled with envisioned bandwidth hungry applications, have triggered the need for high data rates and extremely low communication latencies. Now that dust around 4G has settled down to a great extent, the community is looking for newer technologies to achieve the envisioned requirements of future mobile networks, also referred to as 5G, aiming at improving the stakeholders objective functions. For this concept to be realized, the path is paved by the following technical and economic evolutions: widely available broadband Internet, reduced connectivity costs, more devices being manufactured with built-in sensors and WiFi capabilities, and immense market penetration of smart phones. We believe that 5G - in contrast to earlier generations - will not only improve the end user services, but also go beyond enabling communication between people by realizing machine-to-machine communication and the concept of Internet-of-Things (IoT). The challenges stemming from realizing the vision of *connectivity everywhere* for highly dynamic device layer entities and immensely heterogeneous applications define a major portion of 5G. The communication requirements are even more stringent when the users or sensors are mobile. The expanding networks, the inclusion of many entrants and their dynamic relationships, and virtualized network sections result in a very complex management task of the network. SDN and autonomic network management are seen as the enabling concepts that come to rescue and help in solving the pressing challenges. The SDN architecture propagates the separation of the control plane from the data plane, which enables the flexible hosting of network control functions in different settings and platforms e.g., in centralized or distributed fashions, in physical machines or virtual instances in a cloud network. When applying SDN solutions to the wireless access and mobile networks, it promises to provide efficient management and control over wireless operations by providing a unified management and control platform. Yet, the abstraction of centralized control on one hand and simple forwarding elements on the other, can not be achieved easily in wireless networks. The current SDN technology mostly caters to the needs of wired networks in data center or enterprise network settings and needs to be advanced to allow for monitoring of wireless parameters and for exerting control on wireless network infrastructure within the SDN control layer. A number of approaches to SDNized wireless networks exist in the research literature (e.g., (Kreutz et al., 2015), (Xia et al., 2015) and references therein). OpenFlow is still in its infancy and is evolving to meet the requirements of wireless communication. The Open Networking Foundation presented a report (McKeown et al., 2008) identifying challenges of future wireless networks and how SDN can be a solution to these issues. Research however struggles to include SDN in wireless networks: Some works rely on mobile node cooperation (Schulz-Zander et al., 2015), others create configuration overhead (Suresh et al., 2012), which leads to more load on all involved nodes during handover. Aetherflow (Yan et al., 2015) and its predecessor TinyNBI (Casey et al., 2014) are frameworks that allow the controller to manage the capabilities of a wireless access point (e.g., mapping of logical ports to physical ports, transmission power), to receive events corresponding to the wireless network, or to gather statistics of wireless ports.

The research community has also been actively ad-

dressing the inherent scalability issue of SDNized networks. (Curtis et al., 2011) proposes to offload statistics gathering and management of micro-flows onto switches, thus reducing delays introduced by controller involvment. DIFANE (Rexford et al., 2011) includes switches in the control-plane. *Authority Switches* are introduced, which store a partition of all rules required to operate the network. Instead of switches sending every cache-miss to the controller, they forward the packet which triggered the miss to an *Authority Switch*, which in turn caches the required rules in the ingress switch. In (Rexford et al., 2011), flooding of packets causing a cache-miss is used to reduce packet-loss during handover of mobile nodes. If a node changes its point of attachement, the edge switch the node was connected to floods packets for the mobile node through the network in order to reach the new acces point. With this approach, the new point of attachment may receive packets sent during handover and can forward them to the mobile node, thus reducing packet loss. However, evaluation of the approach is only done by emulating the network and not in a physical testbed.

When it comes to SDN (wireless) networks, researchers often use network emulators like Mininet to evaluate new approaches to utilizing the advantages of SDN. Even though this approach has the advantage of not having to deal with challenges like setting up and configuring physical networks or creating a suitable network environment to simulate real-world conditions, it still lacks the capability to capture the realistic wireless dynamics. In general, the validation frameworks for most of the approaches are abstracted to higher levels, which we believe does not realistically capture the system dynamics. In this work, we discuss a full scale SDNized wireless LAN testbed that we developed using both virtual and physical network entities. The testbed is targeted to study the performance of different contributions (e.g., mobility management, location management, resource allocation, network selection, etc.) in a more realistic environment which involves both wireless and physical mediums, where the topology thickness may be dynamically adjusted for different settings, and where the network management procedures and protocols are fully implemented similarly to that of real deployment. Hence, we claim that our contributed and developed experimental setup will serve as a fitting validation framework for SDN control of wireless networks.

2 SDNized Wireless LAN Testbed

The testbed is designed and developed to function in different modes, of which the two prominent ones are: i) Controller communication mode - in this mode, the trigger events from the forwarding entities are sent to SDN controller, where the decision of trigger handling is carried out. ii) Inter-entities communication mode - inspired by (Jia, 2015), we propose to handle triggering events in a way different to the SDN paradigm of centralized control, i.e. the events are sent to other forwarding entities, which may take over partial control responsibilities delegated by the controller. In this paper, we focus on providing the details of the former due to space restriction. However, it should be highlighted that the two functional modes add additional capability to the testbed. Solutions which deviate from the classical SDN control principles (where based on the Open-Flow protocol, all the trigger events must be forwarded to controller) may be tested and their performance gains may be measured against those following the classical SDN rules. In what follows next, we provide the details of the developed testbed and its components for controller communication mode. Figure 1 pictorially presents the testbed environment, which comprises the following major components:

2.1 Central Manager

It is responsible for high-level policy definition and visualization. The policies are translated into northbound applications, which are then executed by the SDN controller positioned at this layer.

2.2 Aggregator and Transport Network

This component comprises a mesh of forwarding entities, which are implemented using virtual and physical switches. This component forms topology design and decides the topology thickness for different experiment settings (e.g. experiments to study the impact of topology thickness on the trigger propagation, or flow rules definition in the switches on a path).

2.3 Network Edge

This component comprises a set of edge switches that interface the Access Points (APs). It is worth highlighting here that an AP is connected to a port of an edge switch, whereas the wireless access port of the AP provides connectivity to multiple mobile devices. This provisions the change in per-port handling of OpenFlow i.e., a disconnected mobile device does not imply the disconnection of the edge switch's port to which the AP is connected. Hence, the controller needs to implement efficient host tracking to realize SDNized wireless networks. To achieve this, we implement the *WiFi-Monitor* tool, the details of which will follow later in this section.

2.4 Android Measurement Application

This component is a network traffic generating Android application running on Mobile- and Static Node. It can act as both a sender and a receiver. As sender, the application sends UDP packets to a receiving device with a configurable rate. These packets contain the sending timestamp and a packet id. The receiver stores these received packets. In order to automate experiments, the application is able to switch betwen access points by itself. After the experiment execution, the sender sends a *stop*-packet to inform the receiver of the end of the experiment. As logging from the Android device during an experiment would affect the networks performance, the application generates logs at

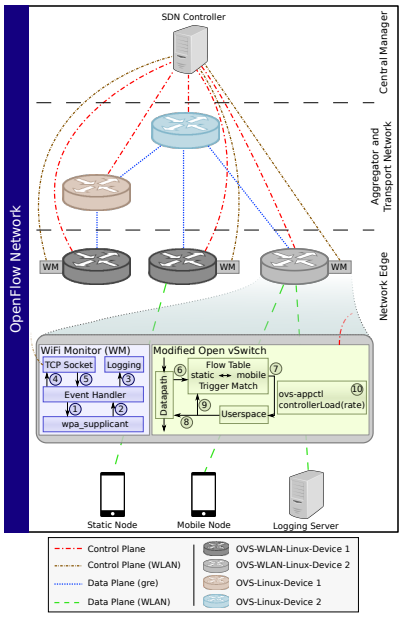


Figure 1. SDNized Wireless LAN - Experimental Setup

the end of experiments.

2.5 Experiment Automation

Running an experiment in a physical testbed is slow compared to an emulated environment. In order to automate both the setup and the execution of the experiment, we develop several tools. The previously discussed Android application is used to trigger the handover process while streaming data.

A machine attached to the network runs the experiment and configures the OpenFlow network using a script. After the setup phase, this machine sends a signal to both Android devices to start the experiment. It is also used to collect logs produced during the experiment from the log server, as well as to present a user interface to show the experiment status and results in real time. We also extended OVS to create load on the controller (configurable via ovs-appctl) by frequently sending dummy *PACKET_IN's* to the controller. As sending this many requests to the controller may have an impact on the switch performance, we use an OVS-Node which is not playing a role in the handover process.

WiFi-Monitor

We develop this tool to achive efficient tracking of mobile hosts required for wireless networks. The tool is implemented using C++ and runs on linux-based systems. As shown in Figure 1, the *WiFi-Monitor* registers itself as a receiver of WiFi-Events to the *wpa_supplicant* module (Malinen) (1). The *wpa_supplicant* is responsible for managing L2-authentication of mobile hosts. As soon as a host connects (disconnects) to (from) an AP, the accounting *WiFi-Monitor* instance receives a callback (2), logs the event (3), and sends a message to the SDN controller application (4). With this message, we forward information about the host (MAC-Address), the event (connect/disconnect), and which AP this event corresponds to (MAC-Address). Using this information, the controller is aware of the position of all hosts in the system at any time after initialization. The *WiFi-Monitor* may also be used for various measurements, of which a few will be detailed in section 3. For these measurements, messages from the SDN controller application are received and logged (3, 5). It should be furthermore noted, that we also modify the OVS implementation to address scalability issues of central control by delegating specific control responsibilities to forwarding entities. However, the implementation details are ommited due to space constraints.

We also test the efficiency of the *Wifi-Monitor* against the Host Tracker service of the OpenDaylight controller (odl, *IfIptoHost*) and found that the built-in OpenDaylight Host Tracker is not well suited to track mobile hosts. If a mobile host changes its point of attachment, the information received from the Host Tracker is mostly outdated and refreshed only after a few seconds. This issue is overcome by the contributed *WiFi-Monitor*. The OpenDaylight application contains a server to which all *Wifi-Monitor* instances connect. On each L2-Event the controller receives from the *WiFi-Monitor* instances, the host-to-AP mapping is updated in order to keep track of the hosts' current positions.

OpenDaylight application

Our ODL application comprises three major parts:

Monitoring: This module contains the server to which the *WiFi-Monitor* instances connect. It internally stores the current Points of Attachement (PoA) of all hosts in the system.

Mobility Learning: The learning module receives a callback from the *Monitoring-Module* on every L2-Event. These events are evaluated in order to predict a host's next PoA. Thereby, a probability vector per host is created $p_h(t,s)$, which describes the probability of a host h to connect to AP s at the discrete time-step t. Each L2-Event triggers an update of this probability vector for the host related to the event. The update step follows the learning framework proposed in (Khan and Tembine, 2012). We avoid further details of the implemented algorithm due to space limitation and partially different focus of this work. However, interested readers are encouraged to refer to it for detailed information. After updating the probability vector and the payoff vector for a host, a new prediction is generated. This prediction estimates the target PoA of the host.

Routing: The *Routing-Module* of our controller application calculates and applies routes to the network. This module uses the *Monitoring-Module* to determine the endpoints (APs) to which the two communicating hosts are attached. Routing can be run in two different modes, namely *reactive* and *proactive*. In *reactive* mode, the routes are only applied when a *PACKET_IN* is received. The *DL_SRC* and *DL_DST* of incoming packets are matched

against installed flow rules, which means source- and/or destination MAC-address of a packet need to match for the flow to be used. When the *Monitoring-Module* receives a L2-Disconnect-Event, all routes for the disconnected host are immediately removed. In *proactive* mode, the controller always installs two routes per connection. For example, consider a host h_2 that has a connection to host h_1. The controller installs two routes in the network and both are associated with h_2. The first route is the active route, which is actively used and contains all flows to connect h_1 with h_2. The second route is inactive. This route is generated using the *Mobility-Learning-Module*. The route contains all flows to connect h_1 (at the current PoA) to h_2 at the predicted PoA. This route has a lower priority than the active route and is therefore never applied by OVS. When h_2 disconnects from its current PoA, the *Routing-Module* will receive a callback (from the *Monitoring-Module*) and all active flows of h_2 are removed, which leads to the inactive routes becoming active. On the next L2-Connection-Event, the behavior depends on the correctness of the last prediction. If the prediction was correct, the connection between h_1 and h_2 is already established. If the prediction was wrong, the installed flows are removed and correct flows are installed. Finally, a new inactive route is installed for the new prediction for future movement of h_2.

3 Mobility Management - A Use-Case Scenario Implementation

Consider the scenario presented in Figure 2, where the mobile user is streaming video from a video server while the mobile device is connected to AP1. The transport network consists of a set of switches. The user now enters the coverage of AP2 and thereby triggers the handover process. We now detail how we realized this scenario and discuss how the handover procedures are carried out in the two modes of developed testbed.

Figure 3 depicts the basic structure of technical components for the use case scenario implementation. All the nodes in the network (with the exception of the hardware switch HWS) are commodity laptops. Where multiple NIC's are required USB-(Ethernet/WiFi) adapters are used. As can be seen in Figure 3, all devices are connected to a switch via Ethernet. The OpenFlow network we use for this scenario, as described in Figure 1, is realized by connecting the OVS-Nodes using GRE Tunnels. Each end-point of the tunnels is seen by the OpenDaylight Controller as a logical port. On the device *OVS-WLAN-Linux-Device 1*, there are two instances of OVS running. Each instance has control over one physical WiFi interface. One of these WiFi interfaces is the built-in NIC, the other WiFi interface is a plugged in USB-WiFi adapter. For simplicity of the scenario, the IP addresses of the mobile nodes are static. This implies that the APs are using the same subnet.

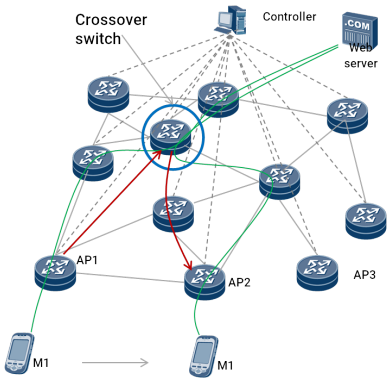


Figure 2. Use case scenario

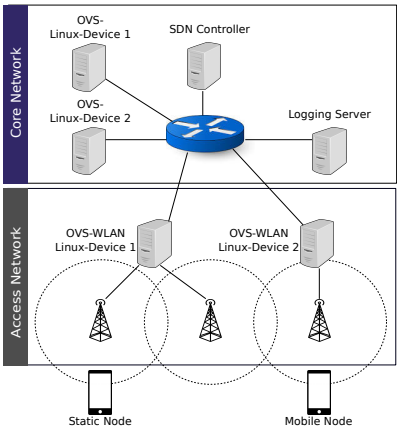


Figure 3. Physical Testbed Setup

3.1 Mobility Handling in Developed Testbed

Upon generation of the first handover trigger, i.e. host M1 disconnects from AP1, the edge switch running *WiFi-Monitor* generates a Disconnect-Event that is forwarded to the SDN-Control-Application. The Application reacts to this event by replying to the *WiFi-Monitor* and by deleting all active flows available for M1 while updating its internal database to store the hosts PoAs. Upon generation of the second handover trigger, i.e. host M1 connects to AP2, the edge switch generates a corresponding Connect-Event, which is forwarded to the SDN-Control-Application, which handles this event depending on the current mode (*reative/proactive*), as described in Section 2.

In this experiment, we capture different delay components, which impact the overall handover cost. To measure the Propagation-Delay (the time the trigger event takes to reach the controller over the transport network) of a layer 2 event, the *WiFi-Monitor* starts a timer directly before sending an Event-Message (connect/disconnect) to the SDN-control-Application, which immediately answers to this event upon reception. The previously started timer is stopped as soon as the *WiFi-Monitor* receives the answer from the controller.

The process to measure L2-Delays is similar. All devices that run an instance of the *WiFi-Monitor* (i.e., ev-

ery AP) are synchronized using the NTP protocol (D.L. Mills, 1985). Every L2-Event is logged to our logging server. The difference in the timestamps of an L2-Event-Pair (i.e., Disconnect-Event and corresponding Connect-Event) is considered to be the L2-Delay.

The Pre-Compuation-Delay is measured within the SDN-Controller and describes how long the reception of the *PACKET_IN* and the handling of it are apart. It describes how long a *PACKET_IN* was queued before being processed.

The Compuation-Delay is also captured within the SDN-Controller and indicates how much time the Contoller needed for calculating the new routes and flows that are created due to a *PACKET_IN*.

The Flow-Setup-Delay is the counterpart to the Propagation-Delay. It describes the time between sending a *FLOW_MOD* from the SDN-Controller and applying it in the OVS. When the Controller sends a *FLOW_MOD*, a timer is started. Each OVS immediately answers to a *FLOW_MOD*. When the Controller receives this answer, the corresponding timer is stopped.

The Flow-Application-Delay is captured within the OVS and describes the time between the modification/insertion of a new flow and the first application of it.

3.1.1 Logging

We utilize the widely used logging module *rsyslog* (logging) for remote logging, which provides the option to forward the system logs to a remote server. We use this mechanism to forward log messages (i.e. previously described delay measurements) from all involved devices to our logging server, which is attached to our local network via Ethernet. Additionally, the logging server connects to one of the APs to be available from within the OpenFlow network. In the SDN-Control-Application, we realize special handling of the logging server. If the *Monitoring-Module* recognizes the MAC-Address of the logging server, an action is triggered to install routes from all APs to the logging server solely using the *DL_DST* field as a match. This way, the logging server is at all times reachable within the OpenFlow network. To realize the forwarding of the Android logs, we utilize the Android Application *Logcat to UDP* (Madzik). We configure *Logcat to UDP* to forward the logs of our Android Application to the remote logging server.

3.1.2 Running the Experiment

Having the testbed setup as described above, we carried out the experiments, in which two Android devices are used to measure the network's behavior during handover. Static Node (as given in Figure 3) is connected to WiFi interface 1 of *OVS-WLAN-Linux-Device 1*. Mobile Node is switching between the two other available WiFi networks. When the experiment starts, Static Node acts as a sender while Mobile Node runs in receiving mode. After Static Node starts sending, Mobile Node switches its PoA several times.

As we know that as the Android device associates to or disassociates from an Access Point, the new topology of the network requires changes to flow tables in switches. These changes are applied differently depending on controller mode. For instance, in reactive mode, the controller installs flows after the first *PACKET_IN* from the OVS-Node at the Access Point from which Mobile Node disconnected. The message to the controller contains the header of the received packet and, depending on the configuration, the packet content. The controller receives the *PACKET_IN* and calculates the path through the network and installs the flows in switches along that path.

However, if the controller is running in proactive mode, the controller predicts to which Access Point a mobile host will connect to in the future. This allows the controller to preconfigure the network with a second route from Static Node to the future location of Mobile Node. This route is installed with a lower priority than the actual route so that only the actual route is used. When Mobile Node disconnects from its current Access Point, the WiFi-Monitor detects the disassociation and sends the information about the topology change to the controller. The controller reacts by deleting the current route, resulting in the 2nd preinstalled route to be the active one. It should be highlighted that proactive controller mode reduces handover delay especially if the disassociation event is propagated to the controller as quick as possible. The vanilla OpenFlow implementation however does not immediately notice that a device left. Together with the *WiFi-Monitor* detecting (dis-)association events much faster, proactive controller mode can significantly reduce handover delay.

3.1.3 Parsing and aggregating results

Each run of the experiment generates log files for each node running software: OVS-Nodes, controller, and Android devices. In order to obtain results, logged events have to be matched to their counterparts and events caused by a single move between Access Points aggregated to evaluate the networks behavior during a single handover. All events but those occuring on the Android devices are immediatly logged to the log server. As we can expect events to have a certain order, only information logged by the Android devices have to be mapped to their respective groups of network events.

The *WiFi-Monitor* is used to improve detection of connection events. We compare the speed and accuracy of Host Tracker to the developed *WiFi-Monitor*. Table 1 shows the statistics of Host Tracker. The *Bad Answers* column describes how often Host Tracker returned a wrong PoA for a queried host. The *Time Lost* column shows how much time was lost until a correct answer was received. This time also depends on the request frequency, which is depending on the amount of *PACKET_IN*s generated. One can see that the behavior is quite unstable with high peaks in the amount of bad answers as well as the lost time. When using the *WiFi-Monitor*, we achieve an average of 9.42*ms* delay over 400 handovers due to the propagation

Table 1. Error Measurement of Host Tracker in 136 Handovers

	Bad Answers	Time Lost(ms)
Average	1.65	121.7
Sum	225	16,556
Max	16	2415

delay. This means that the maximum delay using the Host Tracker (2415*ms*) is higher than the overall delay in 136 runs using the *WiFi-Monitor* ($9.42ms \cdot 136 = 1,281.12ms$). Since the control application reacts to the L2-Events, we do not obtain any bad answers.

4 Conclusions

In this paper, we provided the details of authors' designed and developed testbed for SDNized wireless LAN. We discussed different components of the testbed and elaborated on the technologies used therein. To give better insight of the demonstrator and assist the researchers of SDN topics, we have also provided a discussion on our proposed SDN application and its implementation. The use case scenario summarizes the use of the developed testbed for mobility management.

Acknowledgment

This work is partially funded by iMoveFAN, a collaborative project with Huawei Research Germany.

References

C Jasson Casey, Andrew Sutton, and Alex Sprintson. tinyNBI: Distilling an API from essential openflow abstractions. In *Proceedings of the third workshop on Hot topics in software defined networking*, pages 37–42. ACM, 2014.

Andrew R Curtis, Jeffrey C Mogul, Jean Tourrilhes, Sujata Banerjee, Praveen Yalagandula, and Puneet Sharma. DevoFlow: scaling flow management for high-performance networks. *ACM SIGCOMM Computer Communication Review*, 41(4):254–265, August 2011.

D.L. Mills. Network Time Protocol (NTP). RFC 958, Sept. 1985.

Xiaozhou Jia. SBMP: An SDN-based Mobility Management Protocol to Support Seamless Handover. Master's thesis, The University of Tokyo, 2015.

M. A. Khan and H. Tembine. Random matrix games in wireless networks. In *Global High Tech Congress on Electronics (GHTCE), 2012 IEEE*, pages 81–86, Nov 2012. doi:10.1109/GHTCE.2012.6490129.

Diego Kreutz, Fernando MV Ramos, Paulo Esteves Verissimo, Christian Esteve Rothenberg, Siamak Azodolmolky, and Steve Uhlig. Software-defined networking: A comprehensive survey. *Proceedings of the IEEE*, 103(1):14–76, 2015.

logging. RSYSLOG The Rocket-Fast System For Log Processing. Website. `http://www.rsyslog.com/`; revised July 12, 2016.

J. Madzik. Chemik/logcatudp. Website. `https://github.com/Chemik/logcatudp`; revised July 12, 2016.

J. Malinen. Linux wpa/wpa2/ieee 802.1x supplicant. Website. `https://w1.fi/wpa_supplicant/`; revised July 12, 2016.

Nick McKeown, Tom Anderson, Hari Balakrishnan, Guru Parulkar, Larry Peterson, Jennifer Rexford, Scott Shenker, and Jonathan Turner. Openflow: Enabling innovation in campus networks. *SIGCOMM Comput. Commun. Rev.*, 38(2): 69–74, March 2008. ISSN 0146-4833.

odl. Controller projects' modules/bundles and interfaces. Website. `https://wiki.opendaylight.org/view/Controller_Project's_Modules/Bundles_and_Interfaces`; revised July 12, 2016.

Jennifer Rexford, Michael J Freedman, Minlan Yu, and Jia Wang. Scalable flow-based networking with DIFANE. *ACM SIGCOMM Computer Communication Review*, 41(4):351–362, October 2011.

Julius Schulz-Zander, Carlos Mayer, Bogdan Ciobotaru, Stefan Schmid, and Anja Feldmann. Opensdwn: programmatic control over home and enterprise wifi. In *Proceedings of the 1st ACM SIGCOMM Symposium on Software Defined Networking Research*, page 16. ACM, 2015.

Lalith Suresh, Julius Schulz-Zander, Ruben Merz, Anja Feld- WLANs with Odin. In *Proceedings of the first workshop on Hot topics in software defined networks*, pages 115–120. ACM, 2012.

W. Xia, Y. Wen, C. H. Foh, D. Niyato, and H. Xie. A Survey on Software-Defined Networking. *Communications Surveys Tutorials, IEEE*, 2015. ISSN 1553-877X. doi:10.1109/comst.2014.2330903.

M. Yan, J. Casey, P. Shome, A. Sprintson, and A. Sutton. Aetherflow: Principled wireless support in SDN. In *2015 IEEE 23rd International Conference on Network Protocols (ICNP)*, pages 432–437, Nov 2015. doi:10.1109/ICNP.2015.9.

Information from Centralized Database to Support Local Calculations in Condition Monitoring

Antti Koistinen Esko Juuso

Control engineering, University of Oulu, Finland, {antti.koistinen,esko.juuso}@oulu.fi

Abstract

Maintenance in industry is currently moving from time planned preventive methods to condition-based operation for better process reliability and lowered manufacturing costs. Machine vibrations include information from operating state and machine health and can be used in the computing of several different features for condition monitoring and process control. These describing values can be used for the estimation of remaining useful life (RUL). Local computing enables the use of advanced algorithms for dense vibration data on-site, right next to the monitored process so that the data can be turned into information without the need for large data transfers and centralized computing. Calculated features can be supported with other sensory data, information through expert knowledge, modelling, and data from similar systems in other installations. Developments in wireless technologies enable the use of small nodes in distributed computing. This paper examines the use of locally calculated generalized norms in combination with supporting information from the global maintenance database.

Keywords: intelligent indices, local calculation, edge computing, vibration measurements, generalized norms, combined information

1 Introduction

It has been studied that a large part of the total operating costs in all manufacturing and production plants can consist of maintenance costs. Industry related maintenance costs can vary from 15 percent in food industries to 60 percent in heavy industries of the cost of goods produced. (Mobley, 2002)

This paper introduces advantages of using combined information from several similar targets in addition to just monitor a single target separately. These systems or machines can be located at the same site or at any other location that fits into predetermined criteria. Systems that can be classified to operate in comparable environments make the base for the possible measurement locations. After classification parameters are met for the locations, the valid measurement points can be formed only when the operating parameters for the machinery in these systems match. After all the criteria for valid points are met, these values can be used to improve condition monitoring performance in individual locations. Measurements can be collected along with the meta-data determining measurement conditions and operating parameters and sent to a centralized condition monitoring database. This database provides supporting information to all relevant operators. Information from the database helps in the determining of the threshold level for the amount of stress one machine can withstand, locating different fault characteristics, and improving operating performance through best practices. The determining of the threshold level for machine stress resistance gives the life expectancy for the part and the variation of the measurement points shows the reliability of these results. Operating habits vary between different sites and even within the same site. This framework could include the effects of these different driving habits and reveal the best practices quickly.

Vibration measurements are widely used in industrial applications to monitor condition and operating state of the machinery. Almost all machines vibrate and when the machine operation changes, the vibrations change as well. These changes can indicate shift in machine condition when linked to specific faults. Predicting developing faults leads to minimal down time and better overall control of process maintenance with scheduling and preventing of sudden break downs.(Rao, 1996)

Local calculation enables the use of vast amount of data in condition monitoring and machine control. Advanced feature extraction can be done in small computers located next to the monitored machinery or in the sensor itself. Informative indices extracted from the dense accelerometer data should be used as any other measured data. The applications include long term condition monitoring and determining of remaining useful life that enables the prognostics aspect and real-time operating state detection. These values can be used in control applications, stress monitoring or calculating of condition indices when the machine is operating in the predefined reference state.

Centralized database in a server with versatile interface enables the use of this data in several different locations by varying users at the factory. This local database can be connected to a global framework

providing interoperability and integrability of services (Arrowhead). This work is done in Arrowhead project which develops widely interoperable and integrable service-based collaborative automation framework. Its vision is to enable collaborative automation by networked embedded devices and lead the way to further standardization work. In the following section, short style guidelines are given.

2 Local Calculation

Advances in technology have made the processing of large datasets with small distributed systems possible. Data acquisition (DAQ) system combined with the field programmable gate array (FPGA) can do the data processing while recording it (Shome et al., 2012; Zheng et al., 2014). FPGA core can be faster in certain calculations than comparable digital signal processors (DSPs) and personal computers (PCs) (Vite-Frias et al., 2005). It can be useful e.g. in data pre-processing where it can filter the noise from the vibration signal in real-time (Shome et al., 2012). Small programmable automation controllers (PACs) can be very useful at the algorithm development phase as they can record varying sensory data streams and run calculations for the data. Figure 1 presents the algorithm development for local calculation and generalized third party data usage. The PAC setup that we have used for vibration monitoring cases consists of National Instruments cRIO-9024 controller with cRIO-9114 chassis which has Xilinx Virtex-5 reconfigurable FPGA core. Vibration sensors were connected to NI 9234 analog input module with built in anti-aliasing filter designed for the vibration measurements. Code for data acquisition was developed with Labview software. cRIO can act as a versatile platform for algorithm development for its modular construction and easy configuration.

Determination of the machine state based on vibrations makes efficient maintenance planning possible through predictions of developing machinery condition. It can be also used for planning of machine use in order to prolong its operational time if there is e.g. planned maintenance break coming up.

Vibration data can be used for the automatic maintenance operations. Nowadays, spare parts dealer gets alarm when certain threshold is exceeded and he can react immediately and start necessary preparations for sending replacement parts or planning of repair operation. Data send to third parties from the plantwide database should be carefully secured and only intended parties should have access to this information. Data should be carefully defined with relevant metadata especially for third party users since values without any connection become obsolete. Automation service providers have applications using this presented fast maintenance idea. ABB has a rapid response service (Rapid Response) that promise to provide instant repairs and needed spare parts in an agreed timeframe. They use data from clients machinery to monitor exceptional situations or failures and minimize the process down time. This idea can be further developed by the use of local processing for advanced monitoring methods.

Wireless technologies enable interesting applications for these small devices capable in signal processing. These nodes are capable in data compression and transferring of large amounts of data wirelessly (Huang et al., 2015), data filtering (Ramachandran et al., 2014), and certain transformations (Merendino et al., 2011). Unfortunately nodes have restrictions in measurement accuracy and computing power due to limited battery power and the expectation for the low unit cost. Small sensor nodes can have simple algorithms implemented for filtering or pattern recognition but more complex algorithms would require more processing and thereby more battery power (Ramachandran et al., 2014).

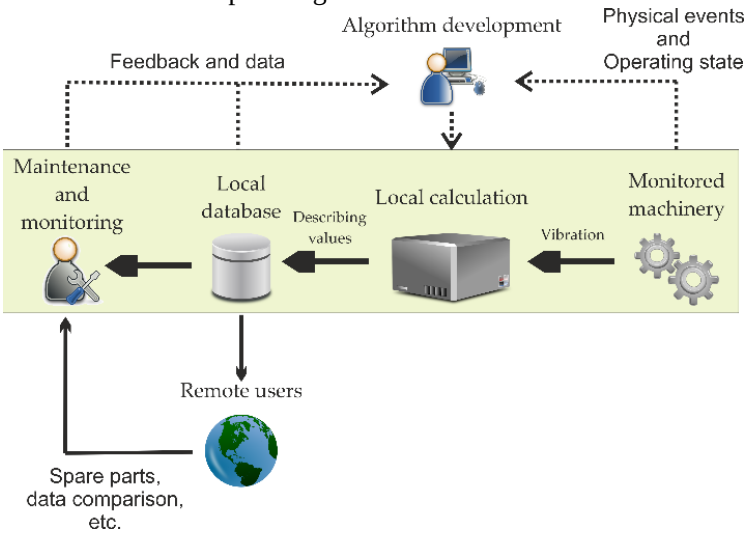


Figure 1. Algorithm development for local calculation and generalized use of extracted features.

In both cases, data transfer of raw measurement data is usually unnecessary and would require high bandwidth. Additionally, in case of wireless sensors the use of energy due to unnecessary data transfer should be avoided (Lahdelma and Juuso, 2011a). Guo and Tse listed several references to available compression methods in (Guo and Tse, 2013) for applications where lots of vibration data needs to be transferred. Huang et al. presented a lossless compression scheme for the wireless sensor network and achieved the average compression ratio of 59.01% (Huang et al., 2015).

3 Signal Processing

Vibration signals can be used in measuring simple vibration severity defined by default as the maximum rms value of the vibration velocities. Peak and rms values are just two common features used in vibration analysis. Vibration signal provides large amount of information and different features indicate different processes in machine operation. Finding the right feature for the wanted event is a matter of referencing the calculated values to machine operation and finding the correlations between these values. Features can be combined to form combined indices which in some cases increase the sensitivity of event detection. Generalized norms can be calculated from the vibration data and used to form intelligent indices using nonlinear scaling.

3.1 Generalized Norms

Vibration data has large amount of information which needs efficient processing. Advanced feature extraction methods can describe large amount of measurement points with one informative value. Generalized norms are described as,

$$\left\|\overline{\chi}^{(\propto)}\right\|_p = \left(\frac{1}{N}\sum_{i=1}^{N}\left|x_i^{(\propto)}\right|^p\right)^{\frac{1}{p}} = \left\|\overline{\chi}^{(\propto)}\right\|_{p,\frac{1}{N}} \quad (1)$$

where, $\alpha \in \Re$ is the order of derivation, $p \quad (1 \le p < \infty)$ is the order of the generalized norm, $N = \tau N_s$ where N_s is the sampling frequency τ is the sample time. Generalized norm is also known as Hölder mean or power mean and it has the same dimensions as the corresponding signal $x^{(\alpha)}$. Some special cases of the norm (1) are arithmetic mean (p = 1), rms (p = 2), and peak value (p = ∞). (Lahdelma and Juuso, 2008a)

Norm calculation compresses five second vibration information of 128000 measurement values (25600 Hz sampling rate) into a single value. Calculation can select e.g. the biggest norm value out of five consecutive values using a sliding window.

Fault detection of fast impact like events can be increased by using derivation of acceleration signal (Lahdelma and Juuso, 2011a). Fault detection has traditionally used displacement $x^{(0)}$, velocity $x^{(1)}$, and acceleration $x^{(2)}$ signals. Higher order derivatives $x^{(3)}$ and $x^{(4)}$ have been previously used in the cavitation detection of Kaplan water turbine (Lahdelma and Juuso, 2008b). Higher order derivatives extend the range of event detection and by selecting correct signal and norm combination, these values can be used widely in different applications (Lahdelma and Juuso, 2011b). Analogue differentiators/integrators can aid in real time calculations (Juuso and Lahdelma, 2006; Lahdelma, 1992, 1995).

Noise from motors and several other mechanisms occurring simultaneously with the monitored property causes false state detection and errors in values. It is important to filter this noise generated by not desired mechanisms before values are calculated. Sensor placing is comparable to the importance of sampling method in manual sampling measurements. Selecting the right order of norms or combination of norms and using high-pass and low-pass filters can be sufficient in most cases. Using of displacement, velocity, or higher order derivatives according to character of the monitored process improves the feature extraction. Different norm values can also be combined to make some events more visible.

3.2 Stress Indices

Stress indices are formed from calculated norms by the means of nonlinear scaling. Norm values are scaled to the linguistic range of [-2, 2] for easy understanding. These scaled values are easy to comprehend and user without deeper understanding about certain measurement from the process can easily use this linguistic range which translates to {*very low*, *low*, *normal*, *high*, *very high*}. These scaled values can be used in decision making and control like any regular process measurements. (Juuso, 2004, 2011a)

Stress indices can reveal sudden high stress areas in machine operation and guide the machine operator or change the customary habits of machine operating cycle. Indices can reveal the remaining useful life (RUL) of the monitored component by summing up indices from more severe vibrations that exceed certain threshold limit. RUL can be estimated when the stress resistance of certain studied part is known. This information can be achieved through monitoring of the part from installation to break down. Figure 2 presents the stress indices and their use in the describing of sudden and cumulative stress.

Stress causes fatigue, which forms micro fractures. This micro fracturing can be seen as rise in the level of stress indices. Indices are scaled according to the current condition of monitored part and the scaling function needs to be updated after the fatigue have caused changes in condition as the old range is no longer valid.

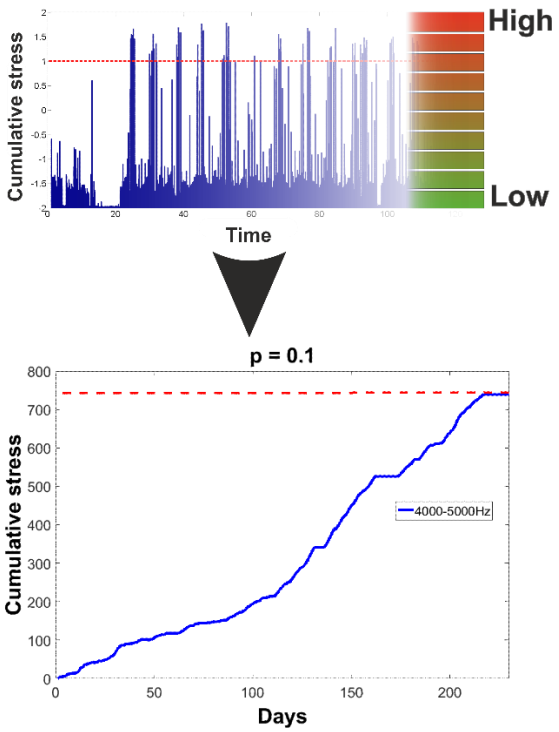


Figure 2. Stress indices scaled to linguistic levels and used to form cumulative stress.

New values can be included in calculations according to changed state and the order of the norms can be re-evaluated if needed.(Juuso, 2011b) Cumulative stress is formed by adding the indices exceeding the threshold level of high stress. Linear increase in cumulative stress indicates that the stress cycles are relatively similar and there have not been any dramatic changes in condition. After the material has experienced enough high load cycles, the micro fractures formed by the stress change the vibration levels and this can be seen as increased slope in cumulative stress meaning that there are more indices exceeding the threshold level for the high stress.

3.3 Measurement Index

Norm values can be used also to track relative changes over time in comparable situations with dimensionless measurement index (MIT). (Lahdelma, 1992) This index has been used in rating of the machinery condition and it is defined as,

$$ {}^{\tau}MIT_{\alpha_1,\ldots,\alpha_n}^{p_1,\ldots,p_n} = \frac{1}{n}\sum_{i=1}^{n} b_{\propto_i} \frac{\left\|\overline{x}^{(\propto_i)}\right\|_{p_i}}{\left(\left\|\overline{x}^{(\propto_i)}\right\|_{p_i}\right)_0} \qquad (2) $$

where norms $\|x^{(\alpha i)}\|_{pi}$ are obtained from the signals $x^{(\alpha i)} = 1,\ldots,n$. The divider represents the state where the machine is in normal operational state, $b_{\alpha i}$ is a weight factor for rating individual faults or events. The sum $\Sigma^{n}_{i=1} b_{\alpha i}=n$ can be combined with other quantities like temperature, pressure, or some statistical features of signals.

Figure 3 presents the use of condition indices in condition monitoring of the load haul dumper front axle. The change in condition can be seen as a strong raise in index level after 250 days.

4 Advanced Wear Monitoring

Remaining useful life can be quite simple to predict if the quality of the monitored parts is similar and the stress constant. Known stress resistance level gives the target value for the probable failure limit and this can be used to predict the expected lifetime rather accurately even without monitoring. The more common case is that the stress levels vary and we need to monitor some

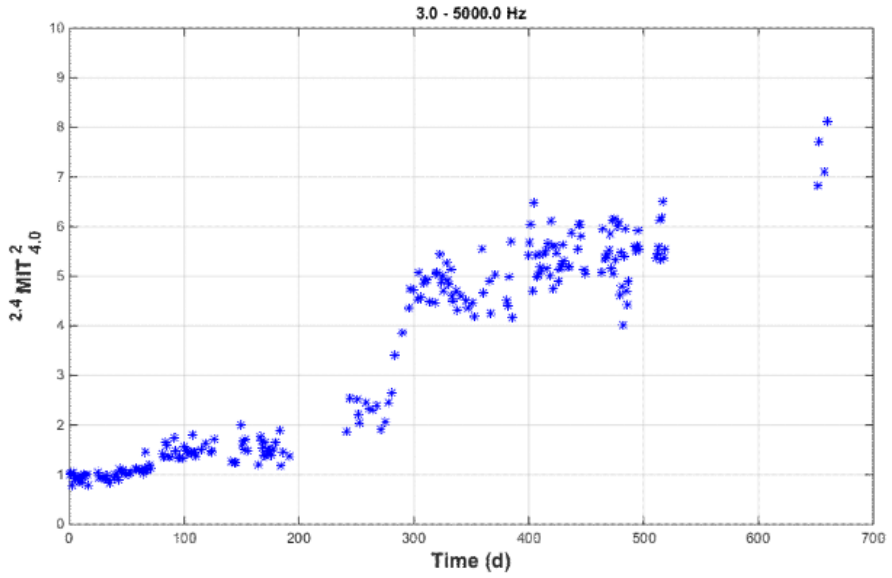


Figure 3. MIT condition indices used in load haul dumper front axle monitoring. (Nissilä et al., 2014)

indicators that tell us about the changes in condition or upcoming failure.

Vibration is a good indicator with rotating or cyclic machinery. The problem here is that the vibrations consist of information from several different mechanisms and we need to filter the data in order to find the valid information. Intelligent indices can isolate the wanted mechanisms of machine operation. These features can then be further used in combination with other indicators in order to strengthen the observations. Increased vibrations with particles in oil or increased temperature can indicate the upcoming failure and this idea can be used with information acquired from other identical setups that have been monitored with similar equipment.

Fault development processes are typically very slow and require long condition monitoring periods. Stress and condition indices require all the information from the installation of new part until the break down occurs to gather the information about the threshold level the part can withstand. This sets high requirements for the monitoring equipment as the locations are not clean and the possibility for cable break or some other failure is high. Single fault gives the data from a single break down and if we want to increase the statistical reliability of the results we need several measurement points. Variation in the results of similar faults gives the probability of break down after certain amount of stress. Characteristics in machine operation and condition monitoring data leading to identified fault can be recorded. Recorded data is now found under this identified fault for building knowledge for the future condition monitoring at all connected sites. Shared condition and stress data makes the determining of RUL more reliable in comparison to monitoring one target alone. It gives various points where the fault has occurred and variation between these points can be used to define probability for the break down if the operation is continued at the same level of stress.

Global condition monitoring database could include the condition information gained with varying algorithms. This requires the scaling of these values into the same universal range (like nonlinear scaling in the forming of stress indices). The database has to use a standardized way of describing data points. Universal descriptions ensure the robustness of the platform and verifies that we are dealing with the right dataset. The database can use a standardized metadata format for making the data exchange as robust as possible. The Open System Architecture for Condition Based Maintenance (OSA-CBM) standardized database of the Machine Information Management Open Systems Alliance (Mimosa) can work as a model for meta-data as it has standardized definitions which help to locate the wanted sensor from the specified machine in certain location (Sreenuch et al., 2013;MIMOSA).

General problem in using shared databases between several operators is the integration to varying systems. Several different clients and languages normally need some proprietary middleware like an application server. Representational State Transfer (REST) uses HTTP methods to transmit data over a wide range of clients written in different languages without the middleware. RESTful API provide data in standardized form according to your data model in flexible way to several different applications. This ensures that all the different operators can use their systems to use the data and provide their own without unnecessary and time consuming changes.(Rodriguez, 2008)

Local signal processing is a vital part in making condition monitoring data into usable form. The database cannot include all the vibration data from every monitored target since the amount of data would be overwhelming and the requirements for the data transfer would be too much. Instead it is reasonable to use feature extraction methods to describe the vibrations with more sparsely recorded values. Figure 2 describes the data reduction that can be achieved by using these feature extraction methods when single describing value is extracted from 5 seconds of raw rod mill vibration data. Raw data is useful to have from situations where machine is working outside of the determined operating state or from some other exceptional situations. This can be done by using triggering for data recording and only save the raw data from exceptional situations since the occasional larger data amounts are not difficult to store.

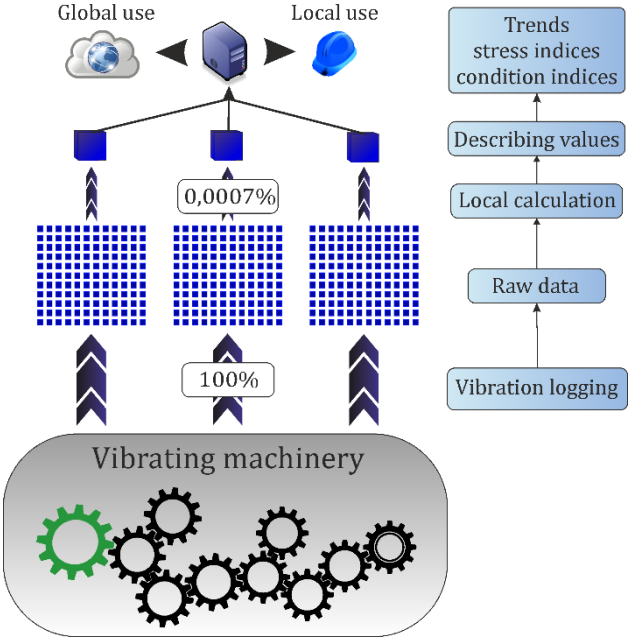


Figure 2. Local calculation in condition monitoring. Data reduction percentages are taken from the calculations done for the acceleration sensor data from rod mill at the Outokumpu Chrome Oy, Kemi mine enrichment plant.

Database information can be used for forming probabilities to back up the local measurements and

indices. They can additionally form different statistical indicators that can be scaled to similar range as the local indices. These additional indices could work like other local measurements from the same machine and give more information to decision making process and maintenance planning. The condition monitoring framework could also act as a gateway to share information about the machine operation and faults. This information sharing could aid the machine and part manufacturers. Manufacturers could shorten the response time to develop more suitable products for specific uses or environments.

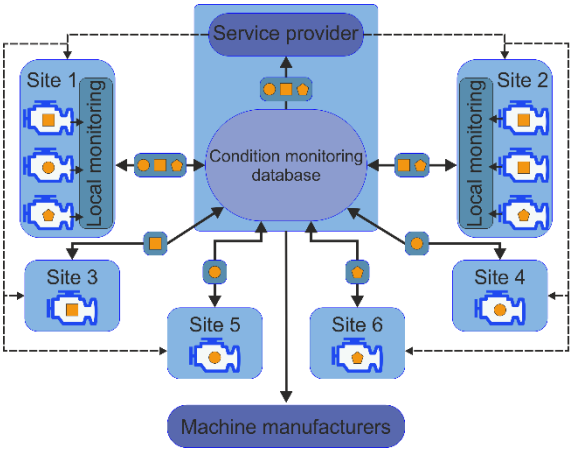


Figure 3. Data sharing with centralized database. Service provider can be e.g. some automation service provider. Orange shapes describe the characteristics of the data.

This idea is not limited to one possible construction only. Figure 5 illustrates the possible framework. Data is defined by its meta-data and data can only be used by the users with privileges so that the data has the pack of users it concerns. Proper certification is needed for this. Maintenance plan of the operator defines its role in this framework. Operator can be both the data provider and the consumer in the case where the monitoring is done at the manufacturing site. Monitoring and analytics can be additionally done by a third party service provider which uses data to develop the operation and to organize maintenance actions. Third party service providers like automation companies have great capabilities to use this data efficiently in their services. The framework would provide important information for the asset lifecycle management and it can help in determining the effects of different factors to asset lifecycle. These effects would also give new ideas to part manufacturers and companies providing machinery.

The Arrowhead framework developed in Arrowhead project can work as a base between different operators sharing their condition monitoring data. The Arrowhead framework is widely interoperable and integrable service-based collaborative automation framework. It visions to enable collaborative automation by networked embedded devices and lead the way for further standardization work. This would enable the service exchange between any actors in the global network. (Arrowhead)

5 Conclusions

Local computing is an effective tool for extracting information from machinery and parts that were earlier impossible due to computational requirements. Localized processing power is relatively cheap in comparison with the savings it can generate through lower down time and improvements in process control. It is inefficient to transfer all the measured raw data to be processed centrally and local computing transforms the data in universally useful and understandable numbers.

The proposed framework takes this locally preprocessed information and makes it useful for several actors. Other operators would benefit from increased information from their processing equipment. Automation and analytics providers could use the information to create new services and add new value to their existing ones. Processing equipment manufacturers would also benefit from increased knowledge about how their products perform at different conditions. Open framework between these operators would enable sustainable development and versatile use of data in several different systems. This is a preliminary work and continuation work includes testing of this idea in practice as a pilot. It also requires further studying in order to find the practical and sound implementation methods.

Acknowledgements

This study was made in the Artemis project “Production and energy system automation and Intelligent-Built (Arrowhead — Ahead of the future)”. Outokumpu Ferrochrome Ltd Kemi Mine is acknowledged for the collaboration.

References

W. Guo and P. W. Tse. A novel signal compression method based on optimal ensemble empirical mode decomposition for bearing vibration signals. *Journal of Sound and Vibration*, 332(2): 423–441, 2013. doi:10.1016/j.jsv.2012.08.017.

Q. Huang, B. Tang and L. Deng. Development of high synchronous acquisition accuracy wireless sensor network for machine vibration monitoring. *Measurement*, 66: 35–44, 2015. doi:10.1016/j.measurement.2015.01.021.

E. K. Juuso. Integration of intelligent systems in development of smart adaptive systems. *International Journal of Approximate Reasoning*, 35(3): 307–337, 2004. doi:10.1016/j.ijar.2003.08.008.

E. K. Juuso. Intelligent Trend Indices in Detecting Changes of Operating Conditions. *In 2011 UkSim 13th International Conference on Computer Modelling and Simulation (UKSim)*, pages 162–167, 2011a. doi:10.1109/UKSIM.2011.39.

E. K. Juuso. Recursive tuning of intelligent controllers of solar collector fields in changing operating conditions. *In Proceedings of the 18th World Congress The International Federation of Automatic Control*, Milano (Italy) August, pages 12282–12288, 2011b. doi:10.3182/20110828-6-IT-1002.03621.

E. K. Juuso and S. Lahdelma. Intelligent cavitation indicator for Kaplan water turbines. *In 19th International Congress on Condition Monitoring and Diagnostic Engineering Management*, pages 849–858, 2006.

S. Lahdelma. *New vibration severity evaluation criteria for condition monitoring*, Research report (University of Oulu), 1992.

S. Lahdelma. *On the higher order derivatives in the laws of motion and their application to an active force generator and to condition monitoring*. D.Sc.Tech. thesis. University of Oulu, 1995.

S. Lahdelma and E. K. Juuso. Signal processing in vibration analysis. *In 5th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies*, pages 867–878, 2008a. doi:10.1109/ISCCSP.2004.1296338.

S. Lahdelma and E. K. Juuso. Signal processing and feature extraction by using real order derivatives and generalised norms. Part 1: Methodology. *International Journal of Condition Monitoring*, 1(2): 46–53, 2011a. doi:10.1784/204764211798303805.

S. Lahdelma and E. K. Juuso. Signal processing and feature extraction by using real order derivatives and generalised norms. Part 2: Applications. *International Journal of Condition Monitoring*, 1(2): 54–66, 2011b. doi:10.1784/204764211798303805.

S. Lahdelma and E. K. Juuso. Vibration analysis of cavitation in Kaplan water turbines. *In Proceedings of the 17th IFAC World Congress*, pages 13420–13425, 2008b. doi:10.3182/20080706-5-KR-1001.02273.

G. Merendino, A. Pieracci, M. Lanzoni and B. Ricco. An embedded system for real time vibration analysis. *In 2011 4th IEEE International Workshop on Advances in Sensors and Interfaces (IWASI)*, pages 6–11, 2011.

K. R. Mobley. *An introduction to predictive maintenance (second edition)*. Burlington: Butterworth-Heinemann. pages 1–22, 2002. doi:10.1016/B978-075067531-4/50000-2.

J. Nissilä, S. Lahdelma and J. Laurila. Condition monitoring of the front axle of a load haul dumper with real order derivatives and generalised norms. *In 11th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies*, pages 407–426, 2014.

V. R. K. Ramachandran, A. S. Ramirez, B. J. van der Zwaag, N. Meratnia and P. Havinga. Energy-efficient on-node signal processing for vibration monitoring. *In 2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, pages 1–6, 2014. doi:10.1109/ISSNIP.2014.6827691.

B. K. N. Rao. *Handbook of Condition Monitoring*. Elsevier. 1996.

A. Rodriguez. *Restful web services: The basics*, IBM DeveloperWorks. 2008.

S. K. Shome, U. Datta and S. R. K. Vadali. FPGA based Signal Prefiltering System for Vibration Analysis of Induction Motor Failure Detection. *Procedia Technology*, 4: 442–448, 2012. doi:10.1016/j.protcy.2012.05.070.

T. Sreenuch, A. Tsourdos and I. K. Jennions. Distributed embedded condition monitoring systems based on OSA-CBM standard. *Computer Standards and Interfaces*, 35(2): 238–246, 2013. doi:10.1016/j.csi.2012.10.002.

J. A. Vite-Frias, R. J. Romero-Troncoso and A. Ordaz-Moreno. VHDL core for 1024-point radix-4 FFT computation. *In International Conference on Reconfigurable Computing and FPGAs ReConFig*, 2005. doi:10.1109/RECONFIG.2005.36.

W. Zheng, R. Liu, M. Zhang, G. Zhuang and T. Yuan. Design of FPGA based high-speed data acquisition and real-time data processing system on J-TEXT tokamak. *Fusion Engineering and Design*, 89: 698–701, 2014. doi:10.1016/j.fusengdes.2014.01.027.

Arrowhead — Ahead of the future. http://www.arrowhead.eu/, 2016.

Rapid response | ABB. http://new.abb.com/uk/service/rapid-response, 2016.

MIMOSA | An Operations and Maintenance Information Open System Alliance. http://www.mimosa.org/mimosa/, 2016.